DOCUMENT RESUME

ED 366 625	TM 021 009

TITLE	The Validity of Testing in Education and Employment.
INSTITUTION	Commission on Civil Rights, Washington, D.C.
PUB DATE	May 93
NOTE	202p.
PUB TYPE	Collected Works - Conference Proceedings (021)
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Achievement Tests; Aptitude Tests; *Civil Rights; *Educational Testing; *Occupational Tests; Personnel Evaluation; Racial Differences; Sex Differences; Sociology; *Standards; Test Bias; Test Construction; Test Reliability; Test Selection; Test Use; *Test Validity
IDENTIFIERS	Standard Setting

ABSTRACT

Because of concerns about the validity of tests used in education and employment, a consultation on June 16, 1989, focused on tests of ability, achievement, and other skills. Invited experts were asked to address a set of issues common to both education and employment testing, primarily related to test construction procedures and how to establish validity. This report contains a background paper identifying key issues, a condensation of the transcript of the consultation, papers written by the panelists, and a summary of their positions with analyses of areas of agreement and disagreement. The papers are: (1) "A Sociological View of Aptitude Tests" (James W. Loewen); (2) "Judging Test Use for Fairness" (Nancy S. Cole); (3) "Bias in Educational and Employment Testing: Selected Issues" (Lloyd Bond); (4) "Standardized Testing: Harmful to Civil Rights" (D. Monty Neill); (5) "A Legal and Policy Perspective" (Clint Bolick); and (6) "Tests Are 'Useful Servants,' Not the 'Masters of Reality'" (Barry L. Goldstein). A major conclusion of this report is that properly designed tests can be used improperly in ways that are unfair and that bias interpretations made of test scores. Standards for test development and use should be set with some form of public involvement. Appendixes provide background on federal guidelines, legislation, and legal issues of test construction. There are three tables and a glossary of testing terms. (Contains 79 references.) (SLD)

****
*   Reproductions supplied by EDRS are the best that can be made   *
*   from the original document.   *
****

TM

ED 366 625

The Validity of Testing in Education and Employment

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

May 1993

A Report of the United States Commission on Civil Rights

TM021009

BEST COPY AVAILABLE

U.S. Commission on Civil Rights

The U.S. Commission on Civil Rights is an independent, bipartisan agency first established by Congress in 1957 and reestablished in 1983. It is directed to:

- Investigate complaints alleging that citizens are being deprived of their right to vote by reason of their race, color, religion, sex, age, handicap, or national origin, or by reason of fraudulent practices;
- Study and collect information concerning legal developments constituting discrimination or a denial of equal protection of the laws under the Constitution because of race, color, religion, sex, age, handicap, or national origin, or in the administration of justice;
- Appraise Federal laws and policies with respect to discrimination or denial of equal protection of the laws because of race, color, religion, sex, age, handicap, or national origin, or in the administration of justice;
- Serve as a national clearinghouse for information in respect to discrimination or denial of equal protection of the laws because of race, color, religion, sex, age, handicap, or national origin;
- Submit reports, findings, and recommendations to the President and Congress.

Members of the Commission

Arthur A. Fletcher, *Chairperson*
Charles Pei Wang, *Vice Chairperson*
William B. Allen*
Carl A. Anderson
Mary Frances Berry
Esther Gonzalez-Arroyo Buckley*
Blandina Cardenas Ramirez*
Russell G. Redenbaugh

* No longer a member of the Commission.

BEST COPY AVAILABLE

The Validity of Testing in Education and Employment

May 1993

A Report of the United States Commission on Civil Rights

TM

Acknowledgments

Eileen E. Rudert directed the project on "The Validity of Testing in Education and Employment," including the June 16, 1989, consultation on the topic and the research and writing of this report. James S. Cunningham supervised the project. Kerry Morgan assisted with the consultation, research into testing applications in the Federal Government, and interpretations of law cases involving tests. Ki-Taek Chun, Harriet Duleep, and Jeffrey O'Connell provided helpful comments on the report. Audrey Wright and Shirley McCoy helped to prepare the report for printing. Editorial assistance and preparation of the report for publication were provided by Gloria Hong Izumi.

Contents

Executive Summary 1
The Report 2
The Panelists 4
Analysis 5
Conclusions 7

Introduction 8
Tests in Elementary Schools 8
Tests for College Admissions and Merit Scholarships 9
Tests for Employment Referrals, Hiring and Promotions 9
Occupation Regulation 11
The Study and Its Scope 11
The Participants 12

Part I: General Issues of Test Validation 15
Definitions of Bias 15
Methods of Test Construction 16
Types of Validity 18
Sources of Test Bias 23
Appropriate Use of Tests 27
Methods of Overcoming Perceived Bias or Adverse Impact 28
Issues 32

Part II: Condensed Transcript of the Consultation 40
Presentation of James W. Loewen 41
Presentation of Nancy S. Cole 45
Presentation of Lloyd Bond 49
Written Statement Provided by FairTest 50
Comments of Alexandra Wigdor 53
Discussion 57
Presentation of Clint Bolick 64
Presentation of Barry L. Goldstein 66
Discussion 69

Part III: Papers by Experts 73
James W. Loewen: A Sociological View of Aptitude Tests 73
Nancy S. Cole: Judging Test Use for Fairness 92
Lloyd Bond: Bias in Educational and Employment Testing: Selected Issues 105
D. Monty Neill: Standardized Testing: Harmful to Civil Rights 118
Clint Bolick: A Legal and Policy Perspective 142

Barry L. Goldstein: Tests are "Useful Servants," Not the "Masters of Reality" 151

Analysis 160
D. Monty Neill, Associate Director, National Center for Fair & Open Testing (FairTest) 160
James W. Loewen, Professor of Sociology, University of Vermont 161
Nancy S. Cole, Executive Vice President, Educational Testing Service 162
Lloyd Bond, Professor, School of Education, University of North Carolina at Greensboro 163
Alexandra K. Wigdor, Study Director, National Research Council, National Academy of Sciences 164
Barry L. Goldstein, Attorney, Saperstein, Mayeda, Larkin & Goldstein 164
Clint Bolick, Director, Landmark Center for Civil Rights 165
Areas of Agreement and Disagreement 166
Conclusions 168

List of Tables

Table 1. Hypothesized Sources of Bias 19
Table 2. Professionals' and Test Developers' Response(s) to Potential Sources of Bias 25
Table 3. Alternatives to Cognitive Ability Tests in Employment 30

Glossary of Testing Terms 170

Table of Select Cases 176

References 177

Appendices

Appendix A: Federal Guidelines and Professional and Agency Standards 181
Appendix B: Major Legislation and Litigation Involving Testing 186
  I. Employment Testing 186
    A. Evolution of Standards for the Use of Tests in Employment 186
    B. Employment Testing in Federal Agencies 187
  II. Testing in Education 188
    A. Elementary Schools 188
    B. Minimum Competency Testing for Students and Teachers 190
    C. Higher Education: Admissions and Scholarships 191
  III. Test Construction Issues—Out of Court Settlements 192

Executive Summary

Tests are used in making a wide range of decisions that affect social mobility and advancement from preschool through employment. They sort workers into jobs and students into schools, classes, and curricula. They often determine who receives rewards, such as college scholarships. Unfortunately, disproportionately few minorities and women appear among those receiving high test scores, a condition referred to as adverse impact. Concern over these test score differences between groups and the frequent litigation over their meaning and fairness prompted the Commission on Civil Rights to undertake a study of the validity of tests and their use in both education and employment.

Four common applications of testing were of particular concern in the study: (1) tests used in elementary and secondary schools; (2) tests used for admissions to higher education and for scholarship awards; (3) tests used for employment referrals, hiring, and promotions; and (4) tests used for regulating occupations. Test score differences between groups have drawn increasing attention to the validity and fairness of the tests in recent years. This attention has resulted in the suspension of tests; the development of new, hopefully more valid tests; and the substitution of certain tests for other tests.

In elementary and secondary schools, many are concerned that tests used for placing students in special classes, for diagnosis of learning disabilities, and for ability grouping or curriculum tracking may unnecessarily segregate students within schools and/or classes and limit their present and/or future learning. The Department of Education's Office for Civil Rights (OCR) can influence test use in schools and education programs funded by the Federal Government and is about to publish rules on the use of tests, particularly for ability grouping.

Meanwhile the new Federal education strategy "America 2000" argues for making schools more accountable by an emphasis on achievement measured by tests and rewarded with scholarships, admission to college, and employment.

Many are concerned that tests used for admissions to college, graduate schools, and technical and professional schools and for scholarship awards determine whether or not and which college students attend. The validity and use of tests has been challenged, sometimes even in court, because scores do not always predict outcomes accurately, are used for purposes other than those for which they were intended, and are often the sole criteria upon which decisions are based. Indeed, in 1989, a Federal judge forced New York State to change its selection process for awarding merit scholarships to high school students because, based on the test alone, girls received lower scores and hence fewer scholarships than boys.

Both the private sector and government use tests for referring candidates to jobs, for hiring them, and for promoting employees. There are three tests used by the Federal Government where challenges to their validity have resulted in recent changes. They are: the Department of Labor's General Aptitude Test Battery (GATB), the U.S. Office of Personnel Management's test for applying to professional level Federal jobs, and the Department of State's Foreign Service Exam.

The Department of Labor's Employment Service administers the GATB to job applicants who are then referred to employers on the basis of their test results. Since 1981, however, the agency has expanded test use to more occupations and experimented with scoring the tests separately within racial/ethnic groups—blacks, Hispanics, and all others—then referring the highest scorers within each race regardless of how they compare across races. The Department of Justice challenged the scoring practice, charging that it constitutes intentional racial discrimination. Uncomfortable with the adverse impact of test scores without the minority group adjustments, the Department of Labor proposed a 2-year moratorium upon the use of the GATB for job referrals while it conducted new studies to improve validity. Before the final directive was issued, Congress passed the Civil Rights Act of 1991, which outlaws the use of race-based score adjustments. Currently, those who use the GATB without the score adjustments have no clear guidance on whether the test will be supported as valid for a broad range of jobs.

On May 22, 1990, the Office of Personal Management (OPM) began administering its newly developed test for applicants to professional level Federal jobs. This test, called the Administrative Careers with America (ACWA), replaces the Professional and Administrative Career Examination (PACE) that was judged to be racially discriminatory in 1982 and streamlines the method for hiring professionals in effect since then. Applicants for Federal jobs in about 100 different series may take the new exam and be hired without agencies evaluating the standard application form (SF 171). In developing the ACWA, OPM strove to achieve merit staffing and a representative work force, and to eliminate adverse impact.

The U.S. State Department suspended use of the Foreign Service Exam (FSE) for recruiting foreign service officers in December 1988. The exam has been the major mechanism by which the Department of State selects the 220 employees it hires annually from among 18,000 to 22,000 interested parties. However, in 1989, a 13-year-old law case charged the Department of State with discrimination against women in its hiring practices. Also, a General Accounting Office (GAO) report to Congress pointed to the oral and written examinations as "barriers that hinder the hiring or advancement of minorities and white women in the Foreign Service." Faced with a test showing adverse impact, and charged with violating Title VII of the Civil Rights Act and an earlier consent decree in the case, the Department of State modified its scoring procedures for examinees in 1988, and suspended further administrations of the test until the concern about adverse impact could be resolved. It is analyzing the skill requirements of the jobs in an effort to redesign the written examination to eliminate any disparate impact.

Federal, State, and local governments and professional associations regulate more than 800 occupations in the United States, including, for example, airplane pilots, cosmetologists, electricians, nurses, pharmacists, physicians, and real estate brokers. Regulation can include licensing, certification, or merely registration. In occupational licensing, for example, the government controls who practices the occupation, typically using examinations aimed at the minimum degree of competency necessary to protect the public health, safety, and welfare.

The use of tests is particularly controversial for certifying teachers. More than 30 States use the National Teachers Examinations (NTE) despite the battery's disproportionate impact on minorities and the shortage of minority teachers. Many States have reexamined their teacher certification requirements and at least one State placed a moratorium on NTE use. The developer of the NTE, the Educational Testing Service (ETS), has promised to replace it with a new battery of tests. The new tests will be available in fall 1993 and are expected to be more valid than the current tests. The current exams rely almost exclusively on a paper-and-pencil format and, according to critics, test only a limited range of minimal competencies—about half of what prospective teachers should know. The new exams will blend pencil-and-paper tests with tests using computer technology, direct observations of classroom performance, portfolios documenting teaching performance, and other items. Furthermore, the tests will be administered three times during a teacher's education and early career, with the final evaluation following a substantial teaching practice.

The Report

Because of concerns about issues such as those above, the U. S. Commission on Civil Rights held a consultation on June 16, 1989, on the validity of testing in education and employment. The consultation focused on tests of ability, achievement, or other skills. Seven experts participated. They were asked to address a set of issues common to both education and employment tests. The issues primarily concern test construction procedures and how to establish validity. This report contains a background paper identifying key issues, a condensation of the transcript of the consultation, papers written by the panelists, and this summary of their positions and analysis of areas of agreement and disagreement. This report is intended to increase knowledge and understanding of how tests are or should be validated, of the controversies that arise in validating tests, and of areas where a consensus may be emerging.

Evidence that racial/ethnic or gender groups respond to tests or their questions differently—that is, that the test questions have different meanings or elicit different answers for different groups—would suggest the test or its items are biased. The primary definition of bias looks to see if test scores consistently over or underpredict performance for members of some subgroup(s). Such a test predicts performance differently for some groups. Known as *differential prediction*, this definition is preferred by most testing experts. It is frequently supplemented with another one. The additional definition looks at group differences in rates of correct responses on test questions or items, making the comparisons among those having the same level of measured ability. However, because ability is usually measured using total test score, average group differences in total test scores are ignored and a systematic bias running through the test cannot be identified. Thus, this second definition is only acceptable if the test has already been validated using the first definition. A third definition is frequently used but unacceptable. It defines bias as group differences in either average test scores or rates of correct responses to test items. With this definition, test score differences could result from other differences between the groups (e.g., in the quality of their education) and their effects would be falsely attributed to the test along with the effects of any test bias.

A test is a sample of questions or tasks intended to provide a quick, efficient, and objective means of drawing inferences about performance. The procedures of test construction are intended to ensure that the inferences are correct. Test developers must ensure that tests are taken in similar environments, have appropriate score distributions and ranges, measure phenomena that are relatively stable over time, produce consistent results if taken again and the measured trait has not changed, and are properly validated.

A test has validity if its scores mean what they should mean. Validation is the process of evaluating the meaningfulness of test scores. External validation establishes the relationship of test scores to other factors; i.e., that the test correctly predicts performance. Such studies are useful for finding systematic biases that run throughout the test. Whether or not systematic biases can be identified and removed from tests may hinge on the appropriateness of the measure of performance and the degree of relationship between test scores and performance.

Internal validation examines the properties of the tests themselves, frequently by examining how different demographic groups perform on the test items. These studies identify test questions that represent extraneous factors, such as bias.

Apart from the two broad types of test validation—external and internal validation—there are several specific types of validity. *Face validity* uses inspection of the test or item to judge whether it measures the intended trait or ability. However, this type is regarded as inadequate to determine that a test is unbiased. *Content validity* is when items are judged as within the relevant content domain or as appropriately balanced with other items according to the frequency of occurrence in the relevant content domain. *Criterion validity* or *predictive validity* is when a statistical analysis shows a systematic relationship between test scores and one or more outcome criteria such that test scores can be used to predict performance. *Construct validity* requires content validity, predictive validity, and face validity and, in addition, inferences that relate what the test measures to other factors and phenomena, such as performance. It establishes that test scores relate to the world in expected ways. But many admit confusion about just how much must be done to achieve construct validity.

The background paper discusses frequently hypothesized sources of test bias. These include sources arising from the test itself, from the test takers (e.g., motivation or test sophistication), from the test environment (e.g., race of the examiner or time limits), and from procedures of test construction and use. Research findings suggest that some biases exist, although they are often for special subgroups (e.g., for those who have not been tested before or recently), or for certain types of test items (e.g., among Hispanics, English words that are false cognates of Spanish words). It is often difficult to decide whether factors that result in test score differences legitimately reflect what the test measures or produce bias.

To minimize potential bias, professionals and test developers use a variety of test construction procedures, instructions to test takers and test administrators, and professional standards and monitoring. They are also concerned about the inappropriate use of tests, such as *overinterpretation* (i.e., using validated tests for purposes other than those for which they are validated). Finally, many methods have been proposed to overcome perceived test bias or adverse impact. They include banning tests, using alternative criteria for selection, emphasizing multiple skills, attributes and abilities, and several others. Each method has advantages and disadvantages.

Key Issues. The test construction issues the study addresses pertain to both internal and external validation. Issues concerning the internal validation of tests include: (1) How should test items that are biased be identified? Is it sufficient that an item is more difficult for one group than another, or should comparisons between groups only be made for test takers with the same test score? (2) Should biased items be categorically eliminated from tests, be kept in when they are strongly related to what the test measures, or be balanced with items having an opposite bias? (3) What proportion of test items in current tests is biased? (4) How much does eliminating items identified as biased reduce test score differences between groups?

Issues concerning the external validation of tests are: (1) Is the predictive validity of tests the same for different racial/ethnic and gender groups? (2) How high should correlations of test scores with performance be for a test to be valid? (3) If predictive validity of a test is high and the same across groups, is it also necessary to establish other types of validity (e.g., content validity or job relatedness)? If so, how?

Apart from test construction issues, the study raised many policy and legal issues. Should State or Federal laws and agencies regulate testing? If so, how? A truth in testing movement has proposed, for example, that test developers file information on test development, validity, etc., with a government agency, and publish tests and their results after test administration. But test administrators argue that publishing a test with correct answers would either increase the frequency, and therefore the cost, of test development, or compromise the validity of results.

Should the use of a test be banned for particular groups when they are judged to be biased? Should test scores be adjusted according to racial/ethnic group?

In court, what evidentiary standards are required to prove disparate impact? When does the burden of proof in such cases shift from the plaintiff to the employer? What standard shall be used to establish that a test is a business necessity? The Supreme Court addressed these issues in *Wards Cove v. Atonio*. Their decision was so controversial that the United States Congress recently passed the Civil Rights Act of 1991 to counteract it.[1]

The Panelists

The panelists who addressed these issues represented a broad spectrum of views. Dr. D. Monty Neill is associate director of the National Center for Fair & Open Testing (FairTest). FairTest's goals are to enhance equity and enable access. Mr. Neill believed that tests, as currently constructed and used, create unfair barriers to achieving these goals.

Dr. James W. Loewen, a professor of sociology, argued that differences in test scores emanate from the social structure. Although some differences in social structure, for example, unequal school finance, affect test scores legitimately, he believed they should not be allowed to legitimize group differences in scores nor to direct attention to individualistic solutions rather than to changes in the social structure.

Dr. Nancy S. Cole, executive vice president of the Educational Testing Service, believed that group differences in test scores or test items should trigger concern about possible bias, but are not necessarily a sign of bias. The scores may reflect valid differences in relevant skills or knowledge created by differences in education and opportunities. She believed the public should take action to ensure that students with low scores are getting help to raise their educational performance. Teachers should *not* assume that those with low scores are unable to learn.

---

[1] The act was passed more than 2 years after the consultation herein was held.

Dr. Lloyd Bond, a professor in the school of education at the University of North Carolina, agrees that group differences in test scores are not sufficient for showing bias. He distinguished the concepts of adverse impact and bias. Biased items should be eliminated from tests, but items should not be eliminated simply because they produce adverse impact. He believed that differences in test scores should reflect differences in achievement resulting from instruction and background.

Alexandra K. Wigdor has directed the National Research Council, National Academy of Sciences' studies on testing. She described the results of their study on the Department of Labor's job referral test, the General Aptitude Test Battery (GATB). The study concluded that the GATB makes useful but not perfect predictions; that its validity would hold for a great many jobs; and that within-group score adjustments can be justified because the errors in test score predictions differ for high and low scorers. The study recommended making score adjustments commensurate with the errors so that qualified people in all groups have the same probability of being referred.

The two lawyers addressed the then-recent Supreme Court decisions and the shifting of the burdens of production and proof and evidentiary standards in disparate impact cases.

Barry L. Goldstein, a civil rights attorney now in private practice, believed that selection practices maintain job segregation. He endorsed the use of tests or other screening devices as a business necessity but not as artificial qualifications. He believed the 1971 Supreme Court decision in *Griggs v. Duke Power Co.*, and the ensuing "Uniform Guidelines on Employment Selection Procedures," improved the caliber of employment practices and the number of minorities employed, particularly in better paying jobs. He disputed claims that test score differences affect productivity and endanger the United States' competitive position in the world economy. He was concerned that *Wards Cove* would reverse the progress in civil rights and make fair employment cases too risky for private attorneys to undertake. Under *Wards Cove*, many selection procedures having adverse impact would remain in place simply because the employers did not intend to discriminate.

Clint Bolick, then director of the Landmark Center for Civil Rights, believed that tests, even when not discriminatory, were automatically abandoned or invalidated prior to the recent Supreme Court decisions. The application of adverse impact analysis is needed to uncover hidden discriminatory practices, but he believed it was expanded to hold employers liable for discrimination when individual preferences, qualifications, or accessibility produced innocent disparities between the racial or ethnic composition of the community labor pool and the work force. The burdens of proof made it relatively easy to challenge tests, but nearly impossible to defend them. He believed the *Wards Cove* decision harmonized adverse impact with Congress' intent to permit the use of professionally developed ability tests when such tests are not designed, intended, or used to discriminate.

Analysis

The Commission's consultation on test construction issues and the longer papers supplied by the panelists convey the nature of the controversy. Neill, Loewen,and Goldstein viewed testing as an obstacle to the important goals of enhancing equity and increasing opportunities. Although Cole, Bond, and Bolick also did not want tests to be unfair obstacles to opportunities, they believed that tests were merely an indicator of other inequalities that minorities face, particularly in the education they received. They emphasize the importance of having accurate assessments because of the many different needs that tests fill.

Despite the wide range of views these panelists held, they revealed many areas of agreement. The following section identifies some major areas of agreement and disagreement.

Definitions of Bias and Discrimination. All of the panelists recognized the potential for bias in tests and for the misuse of test scores in ways that are biased and unfair.

Both the testing experts and the attorneys agreed that average group differences in test scores alone are not evidence of bias.

Each of the panelists listed a variety of potential causes of adverse impact. Most named differences in the quality of education.

Internal Validation—Methods for Eliminating Item Bias. All of the panelists agreed that any items that are biased should be eliminated from tests, although what they regarded as "biased" differs.

Experts' judgments of test questions on their face (i.e., face validity), they agreed, are insufficient for eliminating biased items.

The panelists agreed that test validation procedures must examine individual test items for bias using statistical comparisons for relevant groups. They sharply disagreed over which method should be used. Both Loewen and Neill dismissed as useless methods that compare the difficulties of items across racial or ethnic groups among test takers who have similar overall test scores. Panelists who found methods that adjust for overall test score acceptable did not single out any of these methods as more or less adequate than any others, although one panelist preferred more recent approaches.

Once a method has identified items that may be biased, opinions differed on whether or not those items must be eliminated. Although Loewen agreed that items on which groups differ in performance are not necessarily biased, he believed they should be eliminated from tests to enhance equality. Other panelists would not agree to eliminate the items these methods identify, but they may agree that test developers should provide written justification for continuing to include such items.

Extent of Bias in Existing Tests. Allegations that tests are biased may quantify the extent of that bias by the number of test items that are biased or the proportion of group differences in test scores due to bias. According to Bond, even the better statistical procedures may only identify 5 to 10 percent of trial items as potentially biased. However, not all of the items identified by these methods would be considered biased, and those that were would be eliminated from the test.

Attempts to quantify the extent of bias in tests have often focused on the SAT. Despite their different opinions about test bias and adverse impact, both Bond and Loewen concluded that the largest part of group differences on the math section of the SAT are not due to bias. Bias accounts for at most one-third of the black-white difference in math scores. Their conclusions about bias in the verbal section of the test were much less certain, although both seemed to feel that more of the difference in the verbal was due to bias than in the math.

Methods for External Validation. Our testing experts agreed that methods for eliminating item bias may not be effective when systematic biases run through all the items of a test. Thus, collecting information about how test scores relate to some criterion other than the test itself is critical for validation. All would agree that the external criterion should not be just another test.

The panelists disagreed about whether the predictive validity of tests is the same across sexes or racial groups. They also disagreed about whether small correlations between test scores and performance were adequate for validation. However, all felt that something more than predictive validity is required for validation.

Panelists with generally opposing viewpoints agreed that content should be a driving force in validation studies. For example, school curricula or job duties should determine test content in education and employment applications. Cole believed that even items showing adverse impact should be included if they represent appropriate content.

Monitoring of Test Construction and Use. All panelists voiced support for some form of public involvement in setting the standards for test development and use, whether through advisory boards and forums, the courts, or Federal oversight. The suggestion of establishing Federal oversight for the testing industry, notably, did not draw any strong objections.

All of the experts agreed that properly designed tests can be used inappropriately, but none speculated on how frequently this may occur.

They all agreed that important decisions, such as denial of scholarships, college admissions, or jobs, should not be based solely on test scores. Experience and education or other important selection criteria should be used too.

Mechanisms for Handling Group Differences in Test Scores. The panelists agreed that issues of fairness are separate from issues of bias or adverse impact. They generally agreed that adverse impact will remain in tests even if all bias is removed. However, each proposes a different solution.

Neill suggested doing away with tests in favor of "authentic" assessments such as work samples; at the very least, test scores should be only one of multiple criteria. Loewen recommended removing items showing adverse impact from tests during test construction, even if these items are unbiased. Wigdor and the National Academy of Science's report proposed adjusting test scores for racial/ethnic groups by the amount of error in the test's predictions, so that successful workers in each racial/ethnic group have the same probability of being referred for the job. Her solution was milder than the Employment Service's within-group scoring, which adjusts for the entire difference between groups, but both adjustments are outlawed by the Civil Rights Act of 1991. In discrimination cases, Goldstein would challenge employers to defend all of their selection procedures as essential for the job if any of them shows adverse impact. He would also dismiss the typically low correlations between test scores and performance as too small to validate test use.

In contrast, Cole, Bond, and Bolick thought that tests should be as accurate as possible, regardless of the adverse impact they show. They believed that providing quality education for all groups is the key to eliminating the adverse impact that tests show. Bolick would place the burden of proving discrimination on the plaintiff, lest the employer be held liable for the myriad of innocent causes, such as differences in the quality of education across groups, that may produce adverse impact.

Conclusions

Issues of the validity of employment and education tests continue to arise in Federal, State, and local courts and before Congress. The ways in which tests are used are changing in the Federal Government and in other public and private sectors. The major conclusions of this report are given below.

- Properly designed tests can be used inappropriately, in ways that are unfair and that bias the interpretations made of test scores. Important decisions, such as denial of scholarships, college admissions, or jobs, should not be based solely on test scores.
- Average group differences in test scores do not necessarily reflect bias arising from test construction or use. Differences can arise from bias, which refers to test scores that underestimate the performance of particular groups, and from a variety of other causes, such as differences in the quality of education. Average group differences in test scores may, therefore, remain in tests even if all bias is removed.
- Methods for eliminating item bias may not be effective when systematic biases run through all the items of a test. Therefore, collecting information about how test scores relate to criteria other than the test itself, such as job or school performance, is crucial for validation.
- Biased items should be eliminated from tests. Experts' prima facie judgments of tests questions are not adequate for identifying biased items. Test validation procedures must examine individual test items for bias using comparisons of statistics for relevant groups, for instance by comparing the difficulties of items across racial or ethnic groups among test takers who have similar overall performance. Once a method has identified items as potentially biased, test developers should provide written justification for continuing to include such items in their tests.
- Standards for test development and use should be set with some form of public involvement, whether it is through Federal oversight or public input on advisory boards and forums.

Introduction

Tests are used in making a wide range of decisions that affect social mobility and advancement from preschool through employment. They sort workers into jobs and students into schools, classes, and curricula. They often determine who receives rewards, such as college scholarships. Unfortunately, disproportionately few minorities and women appear among those receiving high test scores, a condition referred to as adverse impact. Concern over these test score differences between groups and the frequent litigation over their meaning and fairness prompted the Commission on Civil Rights to undertake a study of the validity of tests and their use in both education and employment.

Four common applications of testing were of particular concern in the study: (1) tests used in elementary and secondary schools; (2) tests used for admissions to higher education and for scholarship awards; (3) tests used for employment referrals, hiring, and promotions; and (4) tests used for regulating occupations. In the two or more years since the study was undertaken, the validity of tests has been challenged in each of these applications. The challenges have resulted in the suspension of tests, the development of new, hopefully more valid, tests, or the substitution of certain tests for other tests. Some recent developments are given below.

Tests in Elementary and Secondary Schools

Tests are used in elementary and secondary schools for placing students in special classes, for diagnosis of learning disabilities, for ability grouping or curriculum tracking, and for evaluating teachers and schools. The use of tests for pupil assignment raises concerns about possible violations of civil rights. For example, are test scores used to place pupils in ability groups that segregate them within schools?[1] Do pupils with disabilities receive the special classes to which they are entitled or are they placed in classes that limit their present and/or future learning?

The Department of Education's Office for Civil Rights (OCR) has responsibility for enforcing civil rights in schools and education programs funded by the Federal Government[2] and can deny Federal funds to those that do not comply. In practice, their policy on ability grouping[3] requires that some subjects are not grouped by ability and limits the contribution of teacher judgments to decisions about ability grouping. However, OCR has not published any policy or rules[4] on the use of tests for diagnosis and program assignment of students with disabilities and for ability grouping. Such rules could significantly change the ways in which schools use tests.

---

[1] For example, see North Carolina Advisory Committee to the United States Commission on Civil Rights, *In-School Segregation in North Carolina Public Schools*, March 1991.

[2] OCR's jurisdiction falls under Title VI of the 1964 Civil Rights Act, sec. 504 of the Rehabilitation Act of 1973, Title IX of the Educational Amendments of 1972 (concerned with sex discrimination), and the Age Discrimination Act of 1975.

[3] These conclusions are based upon relevant court cases and OCR-initiated administrative proceedings. Two court cases addressing the issue of ability grouping are Montgomery v. Starkville Municipal Separate School District and Quarles v. Oxford Municipal Separate School District. These cases supported the use of achievement grouping when it is used to assist students in their ability to learn and when students are not locked into a given group and can move about between levels.
In enforcing civil rights laws, OCR has, to date, undertaken administrative proceedings against four schools or school districts. (All were initiated in 1984.) In these instances, OCR objected to the rigid use, without educational justification, of composite test scores for ability grouping that resulted in all or predominantly white classrooms. It cited recent research indicating that because students still vary in their mastery of specific subjects, they do not benefit from ability grouping based upon composite scores, but *do* benefit when grouped for specific subjects according to their achievement within that subject. OCR also objected when ability groups were assigned using test scores and teacher judgments based on vague criteria when the classes were even more segregated than if only test scores were used.

Amidst controversy over whether tests are appropriately used in schools, President Bush's education strategy "America 2000"[5] promises to make schools more accountable by establishing "World Class Standards" and voluntary American Achievement Tests, by awarding citations and scholarships based upon test results showing educational excellence, and by encouraging colleges and employers to use the American Achievement Test results. The standards and achievement tests will be developed for each of five core subjects and will represent what young Americans need to know and be able to do to live and work successfully. The educational community's division on national testing and reservations about the fairness to minorities and women of a national test will no doubt affect the strategy as it develops from its proposal stage to implementation.

Tests for College Admissions and Merit Scholarships

Tests are used for admissions to higher education—college, graduate schools, and technical and professional schools—and for scholarship awards. College admissions tests, such as the Scholastic Aptitude Test (SAT) and the American College Test (ACT), show test score differences between males and females and between minorities and whites. In particular, the SAT's stated purpose is to predict how well students will do in their first college year. Many question SAT predictions because girls get better average grades in high school and college than boys, but boys consistently outscore girls on the test. Furthermore, SAT scores are often used for purposes other than predicting first-year college performance. They may determine which college a student attends and whether or not he or she receives a scholarship. The latter use was challenged in court.

On February 4, 1989, a Federal judge in Manhattan ruled that New York State's method of awarding merit scholarships to high school students on the basis of SAT scores discriminated against girls.[6] He ordered the State to change its selection process, although he felt an SAT component was justified.

New York State's effort 2 years earlier to combine SATs with high school grades was abandoned when schools began inflating grades in hopes of having more scholarship winners. In light of the judge's ruling, however, the State has returned to using a combination of grades and SAT scores to award its Empire State Scholarships of Excellence and Regents College Scholarships. Although the State hopes to develop a special test to use, progress has been slow.

Tests for Employment Referrals, Hiring, and Promotions

Tests are used for referring candidates to jobs, for hiring them, and for promoting employees. The private sector and Federal, State, and local governments rely extensively on tests for these purposes. For example, in the Federal Government, the Department of Labor's Employment Service uses ability tests to refer applicants to private sector jobs and the U.S. Office of Personnel Management and the Department of State's Foreign Service use them to hire Federal employces. Some recent developments concerning the validity and use of tests in these three agencies are highlighted below.

---

[4] Rules must first appear in the *Federal Register* as "proposed" and provide a period during which the public may submit their comments. Public comments must be taken into account before the final rules are published. Final rules can assume the force of law when cited by a court.

[5] See: U.S. Department of Education, "America 2000: An Education Strategy" (1991), Rothman (1991) and *Education Daily* (Apr. 19, 1991, pp. 1-3).

[6] See Glaberson (1989), Evangelauf (1989), Holden (1989), Uhlig (1989) and Sharif v. New York State Education Department,—F. Supp.—Feb. 3, 1989 (S.D.N.Y.).

The Department of Labor's General Aptitude Test Battery (GATB). The Department of Labor's Employment Service administers the GATB to job applicants who are then referred to employers on the basis of their test results. Since 1981, however, the agency has expanded test use to many more occupations and experimented with scoring the tests separately within racial/ethnic groups—blacks, Hispanics, and all others—then referring the highest scorers within each race regardless of how they compare across races. The Department of Justice challenged the scoring practice, charging that it constitutes intentional racial discrimination. Uncomfortable with the consequences of using the test without the minority group adjustments to scores, the Department of Labor proposed a 2-year moratorium on the use of the GATB for job referrals while it conducted new studies to improve validity. Before the final directive was issued, Congress passed the Civil Rights Act of 1991, which outlaws the use of race-based score adjustments. Currently, those who continue to use the GATB (i.e., without the score adjustments) have no clear guidance on whether the test will be supported as valid for a broad range of jobs.

The U.S. Office of Personnel Management's new test.[7] On May 22, 1990, the Office of Personal Management (OPM) began administering its newly developed test for applicants to professional level Federal jobs. This test, called the Administrative Careers with America (ACWA), replaces the Professional and Administrative Career Examination (PACE) that was judged to be racially discriminatory in 1982 and streamlines the method for hiring professionals in effect since then. In about 100 different series, applicants for Federal jobs may take the new exam and be hired without agencies evaluating the standard government application form (SF171). The instrument combines a multiple-choice test of reasoning ability and an Individual Achievement Record (IAR). The former measures the ability to understand language, to use reasoning in the context of language, and (except jobs in writing and public information) to solve quantitative problems and problems presented in tabular form. The IAR is a multiple-choice questionnaire about experiences, skills, and achievements in school, employment, and other activities.

In developing the ACWA, OPM strove to achieve merit staffing and a representative work force, and to eliminate adverse impact. OPM staff think the new test will greatly reduce the adverse impact for two reasons: (1) answers to the logical problems of the abstract reasoning portion can be inferred from information provided in the test and use only general knowledge pertinent to the jobs; and (2) educational background and work experience included in the IAR typically shows less adverse impact than tests of abstract reasoning ability.

The Department of State's Foreign Service Exam.[8] The U. S. State Department suspended use of the Foreign Service Exam (FSE) for recruiting foreign service officers in December 1988. The exam has existed since 1924 and has been the major mechanism by which the Department of State selects the 220 employees it hires annually from among 18,000 to 22,000 interested parties. But in March 1989, it was challenged in a 13-year-old law case that charged the Department of State with discrimination against women in its hiring practices. Also, a General Accounting Office (GAO) report to Congress pointed to the oral and written examinations as "barriers that hinder the hiring or advancement of minorities and . . . women in the Foreign Service."[9] Faced with a test showing adverse impact, and charged with violating Title VII of the Civil

---

[7] This information is based upon a June 28, 1990, briefing from OPM staff. Also see Vukelich (1989; and Havemann (1988, 1990).

[8] This information is based upon a Sept. 14, 1989 briefing with officials from the State Department. Also see Gonzales (1989), Purnell (1989), and Palmer et al. v. Shultz, 616 F. Supp. 1540 (D.D.C. 1985); and Palmer et al. v. Shultz, 815 F.2d 84 (D.C. Cir. 1987).

[9] The report was titled "Minorities and Women are Underrepresented in the Foreign Service." It was in response to the For-

Rights Act and an earlier consent decree in the case, the Department of State modified its scoring procedures for those who took the exam in 1988, and suspended further administrations of the test until the concern about adverse impact could be resolved.

While use of the test is suspended, the State Department is analyzing the skill requirements of the jobs in an effort to redesign the written examination to eliminate any disparate impact.

Occupational Regulation

In addition to the ways in which counselors or employers use tests to refer, hire, or promote job candidates, government agencies and professional associations use tests to regulate who practices certain occupations. Shimberg (1982) estimates that approximately 800 occupations in the United States are regulated by States. Others are subject to Federal or local regulation. Regulation can include licensing, certification, or merely registration. In occupational licensing, for example, the government controls who practices the occupation, usually with an exam. The typical licensing exam fails examinees who do not have at least the minimum degree of competency necessary to protect the public health, safety, and welfare. Among the controlled occupations are airplane pilots, cosmetologists, electricians, nurses, pharmacists, physicians, real estate brokers, and school teachers.

The use of tests is probably most controversial in the teaching profession, perhaps because a single test battery enjoys widespread use. The National Teachers Examinations (NTE) are currently used by more than 30 States despite the shortage of minority teachers and the battery's disproportionate impact on minorities.[10] Many States have reexamined their teacher certification requirements, and at least one State placed a moratorium on NTE use.

In response, the Educational Testing Service (ETS) has promised to replace the NTE with a new battery of tests.[11] The new tests will be available for use in the fall of 1992 and are expected to be more valid than the current tests. The current exams rely almost exclusively on paper-and-pencil technology and, according to critics, test only a limited range of minimal competencies—about half of what prospective teachers should know. The new exams may use pencil-and-paper tests, too, but will blend tests using computer technology with direct observations of classroom performance, portfolios with documentation of teaching performance, and other items. Furthermore, the tests will be administered three times during a teacher's education and early career: (1) during the sophomore year to evaluate basic skills; (2) at the end of their teacher-education program to evaluate their knowledge of subject matter and the principles of teaching and learning; and (3) following a substantial teaching practice to evaluate classroom performance. The goal is to measure the essence of teaching—problem-solving, decisionmaking, and management techniques that produce effective classroom performance.

The Study and Its Scope

The applications show that validity is indeed frequently at the heart of the controversy over test use. For this reason the study focuses on what is, can, or should be done to validate tests. Thus, the issues address test construction or administration procedures and test scoring. Basic knowledge about how tests are constructed, what bias looks like and how it is minimized, and what makes a test valid informs the policy issues of whether or not tests are appropriate or fair in each of these applications.

The study is primarily focused on cognitive tests, which are mental tests consisting of items based on performances that can be objectively

---

eign Relations Authorization Act, Fiscal Years 1988 and 1989, which directed GAO to review the Foreign Service merit personnel system.

[10] See, for example, Goldstein (1987), Fields (1988a and b), Bradley (1990), and *Professional Regulation News* (October 1989, p. 2).

[11] See Fiske (1988), Watkins (1988), *Education Daily* (Mar. 29, 1990, p. 4).

scored as right or wrong, better or poorer. Included are intelligence tests (tests of abstract reasoning), achievement tests (tests of acquired knowledge), and aptitude tests (tests of special skills or abilities, including intelligence). Generally, cognitive tests are paper-and-pencil tests, but cognitive tests that are oral or administered by video or computers fall within the domain of the study. Tests of skills that are job-related but not necessarily cognitive (e.g., typing) are also included.

Honesty tests and drug or medical tests are sometimes used in employment screening, but the issues they raise are somewhat different from those that arise with tests of knowledge, skills, and abilities. Thus, they are not part of this study.

The study has focused upon testing within the normal range of performance. The available resources could not cover the use of tests, appropriate accommodations, and interpretation of scores for persons with disabilities, which is a subject worthy of study in itself.

This report contains a background paper that provides a common understanding and identifies the key issues for the study, a condensed transcript of a consultation held June 16, 1989, papers from six professionals in the area of testing, and a brief analysis and summary of areas of agreement and disagreement among the experts. It also includes appendices describing Federal guidelines and professional and agency standards used to protect test takers and ensure the quality and fairness of tests and major legislation and litigation involving tests. A glossary defines terms used in the field of testing.

The Participants

Participants in the study included the six expert panelists, a guest speaker, the Commissioners, and Commission staff. The experts were invited to prepare papers and participate in a consultation. The guest speaker, Alexandra Wigdor, was invited to the consultation to present the findings of the then-newly released, government-funded report, *Fairness in Employment Testing*. Biographies follow.

Clint Bolick is a frequent speaker and publisher of books and articles on civil rights issues and legal and policy aspects of testing. An article of his was published in a special issue of the *Journal of Vocational Behavior* (December 1988) devoted to fairness in employment testing. He has a J.D. from the University of California, Davis. He has been special assistant to the Assistant Attorney General, U.S. Department of Justice, Civil Rights Division, and to the Vice Chairperson, Equal Employment Opportunity Commission (EEOC). At the time of the consultation, he had recently become the director of the Landmark Center for Civil Rights in Washington, D.C. His foundation was representing parents of a black student who was not allowed by the California education department to take an IQ test in *Crawford v. Honig*.

Dr. Lloyd Bond is currently on the faculty of the School of Education, University of North Carolina at Greensboro. He earned his Ph.D. in psychology at the Johns Hopkins University and was affiliated with the learning research and development center at the University of Pittsburgh for some time prior to moving to North Carolina. He is recently retired from the board of trustees of the College Board.

Dr. Bond's research has analyzed the thought processes of black and disadvantaged respondents to SAT questions to understand why they are unable to give correct answers. He has also published a number of articles on testing validity and spoken on many occasions, including at a hearing on the effects of testing on black Americans sponsored by the National Commission on Testing and Public Policy (December 1988).

Dr. Nancy S. Cole, from Princeton, New Jersey, represents both herself and the Educational Testing Service.

The Educational Testing Service (ETS) is a private, nonprofit corporation devoted to measurement and research, primarily in the field of education. It was founded in 1947 by the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board. Today it has an annual budget of about $160 million and employs more than 2,000 people.

ETS is best known for developing and administering the College Board's Scholastic Aptitude Test (SAT), taken by about 1.5 million college-bound high school juniors and seniors each year. ETS performs the same functions for many other academic and employment testing programs. Results of their tests are used for school and college admission, student guidance and placement, awarding degree credit for independent or advanced learning, occupational and professional licensing and certification, and awarding continuing education units. They also administer more than 100 scholarship programs and conduct student financial aid services, analyzing applicants' financial needs and reporting them to institutions and agencies, to help distribute available grant and loan funds.

Most ETS testing programs are conducted under contract with independent agencies or organizations. These external groups sponsor the programs, set policy, and determine the overall content of the test.

Dr. Nancy S. Cole became executive vice president of ETS in April 1989 after serving for 4 years as dean of education and professor at the University of Illinois at Urbana-Champaign. Before that she worked at the American College Testing Program and at the University of Pittsburgh. She is a scholar in the field of educational measurement, focusing specifically on issues of test bias, the measurement of vocational interests, and the testing of educational achievement. She has served on the Graduate Record Examination Board, the Committee on Psychological Tests and Assessments of the American Psychological Association, and as president of the National Council on Measurement in Education.

Barry L. Goldstein was an assistant counsel for the NAACP Legal Defense and Educational Fund in Washington, D.C., at the time of the consultation. Since then, he has entered private practice with the firm of Saperstein, Mayeda, Larkin & Goldstein in Oakland, California. Mr. Goldstein graduated from Columbia Law School and received a post-law school degree from the University of Cambridge. During his 18 years with the Legal Defense Fund, he litigated many employment discrimination and other civil rights cases in the district and appellate courts and in the Supreme Court. He was counsel in *Albemarle Paper Company v. Moody*, the first case to rely upon the EEOC Uniform Guidelines for standards of test validation.

He is a frequent lecturer on employment and civil rights law and litigation procedures. In 1985 he was a lecturer in law at Harvard Law School where he taught a course in employment discrimination law. He has also spoken about employment testing issues at conferences such as the Bureau of National Affairs' 1988 conference on that topic.

Mr. Goldstein is cochairman of the Equal Employment Opportunity Committee of the Labor and Employment Law Section of the American Bar Association.

Dr. James W. Loewen holds a degree in sociology from Harvard University. As an associate professor he taught at predominantly black Tougaloo College for 7 years, then moved to the University of Vermont where he is now professor of sociology. In 1990–1991, he was a Smithsonian Fellow in Washington, D.C. Dr. Loewen was also a Fulbright Professor at LaTrobe University in Australia.

He has testified in many court cases and wrote a book, *Social Science in the Courtroom*, describing how to be a legal expert, especially in civil rights cases. He has also written about the race, gender and rural/urban bias in Scholastic Aptitude Test items. He was a speaker at FairTest's 1988 National Testing Reform Conference.

Dr. D. Monty Neill is associate director of FairTest, National Center for Fair & Open Testing, located in Cambridge, Massachusetts. FairTest is a national research and advocacy organization working to ensure that the several hundred million standardized exams administered annually to America's students and job applicants are fair, open, and educationally sound. It was formed in 1985 and works with standardized testing in three areas: elementary and secondary schooling, university admissions, and professional licensing and employment. It is concerned about the overuse and misuse of standardized tests; the harmful effects of testing on individuals, education, and society; and the existence of race, gender, and class biases in testing. In addition to promoting fair testing, it supports the use of alternative methods of assessment.

Alexandra Wigdor participated in the consultation as a guest speaker. She is coauthor of *Fairness in Employment Testing*, a report the National Research Council of the National Academy of Sciences prepared on the U. S. Department of Labor's job referral test, the General

Aptitude Test Battery (GATB). The report was released approximately 2 weeks before the consultation.

Ms. Wigdor has written about the policy issues of test use for many years. She is the co-editor of earlier work published by her organization, *Ability Testing: Uses Consequences, and Controversies*, volumes I and II.

The Commissioners present for the consultation included then-Vice Chairman Murray Friedman, Esther Gonzalez-Arroyo Buckley, Sherwin T.S. Chan, Robert A. Destro, Francis S. Guess, and Blandina Cardenas Ramirez. Commissioner William Barclay Allen, then Chairman, and Commissioner Mary Frances Berry were unable to be present for the consultation. Commission staff who took part in the consultation included Melvin L. Jenkins, then-Acting Staff Director, James S. Cunningham, then-Assistant Staff Director for the Office of Programs, Policy, and Research, and Eileen E. Rudert, project director.

Part I

General Issues of Test Validation

The existence of differences in average test scores between blacks and other minority groups and whites and, on some tests, between males and females is widely recognized. However, there is disagreement over whether test score differences relate to the underlying abilities that tests attempt to measure, and hence to the performance that tests are used to predict, or are merely artifacts produced by irrelevant disturbances. If tests are biased, an obvious remedy is to remove the bias. However, if tests are not biased, further disagreements arise over if and how the adverse impact of tests should be eliminated.

Definitions of Bias

Test bias commonly refers to differences in test scores unrelated to the performance the test is intended to measure. Test developers examine tests and their questions for evidence that racial/ethnic or gender groups respond to them differently; in other words, the questions may have different meanings for different groups. Thus, researchers form hypotheses about how bias might be manifested in test results and look at group differences in test scores and answers for the patterns that bias might be expected to show. The expected patterns of bias are represented in mathematical formulae and are identified during test construction with statistical procedures that we refer to as bias detection techniques. The methods of detecting bias do not identify or interpret the source of the bias. Sources of bias must be inferred by piecing together the common themes of various biased items or test types.[1]

Because the methods of test development search for bias in this generic form, the definition of bias is central to the research.

Bias as differential prediction occurs when test scores consistently over or underpredict performance for members of some subgroup(s). The process of test validation addresses whether or not a test predicts academic or job performance. The question of bias is whether or not these predictions differ for various subgroups, such as racial/ethnic or gender groups. Thus, the predictions are analyzed separately for subgroups.

Educational researchers, industrial psychologists, textbook authors, and even test critics endorse and advocate the definition of bias as differential prediction (e.g., see Schmidt and Hunter, 1974: 1; and Wigdor and Garner, 1982a). Differential prediction is always the primary definition of bias, although other definitions are sometimes used. Rather than looking at test scores, supplemental definitions of bias examine test questions, where each subpart requiring a response is called an *item*.

Bias as different correct response rates (controlled for total test score). If rates of correct responses on test items differ between groups when making comparisons among those having the same level of ability, the items may be biased. Total test score is typically used to identify comparison groups with the same level of ability.

---

[1] Test developers also attempt to identify bias by looking for potential sources of the bias. This approach is discussed below.

A variety of statistical methods[2] may be used to identify such items. The methods are similar in that:

- They ignore average group differences in total test scores.
- They assume that an analysis of differential prediction has already been done and that the exam is largely free of bias. If so, unbiased items will anchor the statistical results and provide contrast for the biased items. If not, the biased items may not be detected.

Newer methods based upon "item response theory" are much more sensitive and flag many more items than older methods (Shepard, 1987). Item response theory methods of identifying bias, however, often require large numbers of test takers (1,000 or more per group) and are therefore expensive or impractical to apply. Thus, less sensitive methods of identifying bias continue to be used, especially with smaller testing programs and smaller minority groups.

Test developers would not necessarily conclude that questions flagged by any of these techniques (and especially the more sensitive ones) were biased without additional scrutiny. They would examine other characteristics of the item (e.g., how the item relates to the total test score) and look for what all the flagged items have in common (e.g., item type or content).

Bias as mean differences (where "mean" is the statistical term for "average") is when group differences in either average test scores or rates of correct responses to test items are regarded as prima facie evidence of bias. Most psychologists challenge this definition (e.g., Flaugher 1978: 673; Shepard, 1987; Wigdor and Garner, 1982a: 70). Snyderman and Rothman (1988: 111) state that mean differences are "an improper definition [of bias], since by taking the existence of group differences as prima facie evidence of bias one begs the question."

Although unacceptable to most psychologists, the definition of bias as mean differences has been used in court. In issuing his decision on the *Larry P. v. Riles* case, Judge Peckham accused the defendants of "unlawful segregative intent" arising from "an impermissible and insupportable assumption of a higher incidence of mental retardation among blacks."[3] This judgment rejected the scientific evidence that differences in performance exist apart from their manifestation in test scores.

Summary. Test bias is when test scores consistently over or underpredict performance for members of some subgroup compared with test takers in general. This definition, referred to as differential prediction, is the only fully adequate definition of bias. Group differences in rates of correct responses on test items among examinees having the same ability is an acceptable definition of bias only when tests have already been shown to have no differential prediction.

Methods of Test Construction

A test is a sample of questions or tasks. It is a quick, efficient, and objective means of drawing an inference about some relevant performance (e.g., in school or a job). Inferences from test scores typically suggest how the ability of an individual with a particular test score compares with (a) other individuals of similar age or (grade) level (for a norm-referenced test), or with (b) the requirements of a set of tasks (for a criterion-referenced test).

---

[2] *Item response theory* (IRT) and *differential item functioning* (DIF) are the most frequently mentioned theory and an accompanying methodology for determining whether test items are biased. (See the glossary for definitions of terms.) The *Mantel-Haenzel statistic* is one mathematical formula used to test whether test items obtain different responses from groups once IRT and DIF procedures are applied.
Some other technical definitions used to detect bias include: *race-by-item interaction* (a sex-by-item interaction is similarly defined); and *different factor analytic solutions* (see factor analysis).

[3] 495 F.Supp 926 (N.D. Cal. 1979), *aff'd in part, and rev'd in part*, 793 F.2d 969 (9th Cir. 1984).

The procedures of test construction are intended to ensure that the inferences are correct. They include standardization, validation, and studies of reliability and stability.[4] Studies of test and item bias are part of the validation process. A large pool of items, administered to people like those to whom the test will ultimately be given, is the basis for all test construction procedures.[5]

Standardization. Test developers design (norm-referenced) tests to distribute the test-taking population across high and low scores with meaningful distinctions. This process is called "standardization."[6] It involves choosing test questions that range around an appropriate level of difficulty and converting each test taker's number of right answers to a score that expresses the person's standing compared with others of appropriate age or (grade) levels.

Tests are often thought biased when proportionate numbers of blacks and other minorities are not included in the standardization (Snyderman and Rothman, 1988), as happened with IQ tests developed in the 1920s. The failure to include minorities or other groups when developing a test can certainly give rise to test bias, because comparisons between groups cannot be made to eliminate unfair questions. However, these comparisons are made during validation (discussed below). The process of standardization has a different goal (i.e., obtaining test scores that distinguish high and low performance) from that of eliminating bias (i.e., ensuring that high and low performance have the same meaning for all groups). Thus, improper standardization per se is not a source of bias in test score differences. Restandardization may change absolute scores and may provide larger or smaller distinctions between some group members, but will not change the order of individuals' scores.

Some of the concerns about whether minority groups have been represented in the populations used to construct tests are avoided by developing tests with scores referring to the ability to do certain tasks (i.e., *criterion-referenced tests*) instead of to the abilities of other individuals (i.e., *norm-referenced tests*).

Although restandardization does not reduce test bias, test developers should carefully represent minorities in the groups on which tests are standardized so that they can apply the procedures that do reduce bias.

External and internal validation. The procedures that attempt to eliminate bias during test construction are external and internal validation. External validation establishes the relationship of test scores to other factors, usually measures of performance of the sort for which the test is used to base selection decisions. Such studies are useful for finding systematic biases that run throughout the test.

Internal validation examines the properties of the tests themselves, generally using test items rather than total test scores. These studies fine tune tests by identifying items that should be eliminated because they represent extraneous factors such as bias.

In principle, external validity, or the absence of a systematic bias, is of greater importance than refining tests by examining individual questions. In practice, however, the information required to conduct internal validity studies (e.g., high school seniors' responses to test items) is available long before that needed to conduct external validity studies (e.g., the relationship between college freshman grades and high school seniors' test scores). Thus, fine tuning is often done before external validation can be undertaken. However, if the test as a whole is not free of bias, then techniques that look at bias in the items can only be partially effective. Hence, whether or not tests are cleansed of systematic biases is critical.

---

[4] In order to be reliable, a test must produce consistent results when administered again to the same individuals. It must show changes only when the trait or ability that the test measures has changed. Furthermore, if the trait or ability the test measures fluctuates on a day to day or hourly basis, the test may not be very useful over time.
Tests of knowledge, ability, and skills are generally reliable and stable, so these properties are not at issue here.

[5] Exams such as the Scholastic Aptitude Test include an unscored subsection of new items each year to select items for use in future tests. "Truth in Testing" advocates are encouraging legislation to require that tests identify this subsection so those taking the test can skip it, if they so choose.

[6] See also the definition of *standardized test* in the glossary. Discussions about standardization are often ambiguous in whether they refer to ensuring that the tests are administered under uniform conditions or to the statistical procedures that ensure that test scores have meaningful differences (as above). Both are necessary to make correct inferences from tests.

Choosing the criterion. External validation is intended to show that the inferences drawn from the test are correct, that is, that the test correctly predicts performance. First, however, a study must measure performance. How one measures performance is not always clear. If the purpose of a test is to select students who will do well in college, is college performance measured by freshman grades, by whether one graduates in 4 years, or by achievement after the college years? Should nonscholastic accomplishments be included in the assessment? If tests measure typing, filing, and phone answering skills, will that represent the performance of a secretary?

When the measure of performance is improper or irrelevant, bias may result. For example, the measure of performance may represent some unnecessary skills that the test is not intended to measure. However, by validating the test using that measure of performance, the test will incorporate the extraneous skills and penalize any groups who lack them. Bias may also result because the measure of performance is not comprehensive enough. When the measure of performance includes too many or too few knowledges or skills or is otherwise improper or irrelevant, the bias is attributed to using the wrong criterion (see table 1).

Studies of external validity measure the relationship between test scores and performance. Because of the many extraneous factors that affect social and economic phenomena and the tendency to capture either too many or too few skills in the measure of performance, predictions are never perfect. How much of a relationship with performance must a test have to be useful or fair, especially when the tests have adverse impacts on certain groups?

Those who argue about whether or not systematic biases are removed from tests often disagree about the appropriateness of the measure of performance and the degree of relationship between test scores and performance.

The value of a test to a user depends upon its benefits compared with using no test, and its cost[7] relative to other measures of the same phenomenon. Many alternatives to currently used tests have been proposed, but they also require validation and may have only limited value. Issues surrounding the value of tests and their alternatives will be discussed later.

Types of Validity

Although there are two major types of test validation—external and internal validation—there are several specific types of validity—face validity, content validity, criterion validity, predictive validity, and construct validity. Face validity and content validity are more often associated with internal validation because they typically examine test items. Criterion validity and predictive validity, which is the primary type of criterion validity, are methods of external validation. Construct validation is all encompassing and can involve either or both internal and external validation techniques.

Face validity is the appearance that a test (or test item) gives of measuring the trait or ability that it is intended to measure, as judged by inspecting the test (or item). Thus, Judge Grady's examination of test items in the ***PASE v. Hannon***[8] case relied upon face validity to establish bias, after he begged in vain for item analysis (Elliott, 1988). Because opinions often differ on which items are biased, face validity is not regarded as sufficient to determine that a test is unbiased. The Equal Employment Opportunity Commission's (EEOC) Uniform Guidelines on

---

[7] The costs and benefits may be assessed either from the perspective of the test user, for example, an employer, or from that of society.

[8] 506 F. Supp. 831 (D.C. Ill. 1980).

Table 1
Hypothesized Sources of Bias

I. Sources Arising from the Test Itself

Cultural bias is when test items contain information that is specific to the culture of one group and absent, to some degree, from the culture of another group.

Content bias. Tests containing questions with content to which some subgroups of the population are less likely to be exposed could show group differences. This type of bias can arise, not just from cultural differences, but from differences in individual interests and school tracking systems, for example.

Sex bias refers to gender differences in test scores that occur despite both groups having the same degree of relevant skills and abilities. It may also refer to the use of content preferred by one gender rather than the other, gender-specific or sexist language, and sex-role stereotypes.

Language may be a source of bias when knowledge of English is not the skill that is being tested and nonnative English speakers or those who speak nonstandard English have difficulty comprehending test instructions or questions. For native English speakers, a language bias may occur when tests use an unnecessarily high level of language. The bias may result from differences in familiarity with or knowledge of the words and linguistic structures of Standard English.

II. Sources Arising from the Test Takers

A. Motivational, Attitudinal, and Other Personality Factors

Test anxiety is thought to produce extraneous thoughts that interfere with concentration and short-term retention. Thus, individuals or groups with higher levels of test anxiety may not demonstrate their true performance level on tests.

Achievement motivation is a general striving to do one's best in activities that can be judged on excellence. Test differences could therefore result from differences in motivation to do well on the test rather than from differences in ability.

Self-esteem. If individual or group differences in feelings of self-esteem or self-confidence have a greater effect on test scores than on the performance the test is intended to predict, test results would be biased.

Reflection-impulsivity. Reflective persons tend to delay responses in answering test items involving an initial uncertainty. With the additional time spent in answering questions, their performance gains accuracy. Impulsive persons respond quickly, often at the expense of more errors than the same persons would display if their responses could be delayed. Thus differences between groups in this tendency could affect test scores.

B. Test Sophistication

Practice. Practice effects are gains in test scores as a result of taking the test (or a similar form of it) over. They may result from familiarity with test format, limiting the time spent on doubtful or puzzling items, or other forms of test sophistication. Bias would result if some groups perform better on tests because more of them have taken the test, or ones like it, before.

Table 1 (continued)

Access to coaching. Coaching generally includes instructing the test taker in test taking procedures, such as how to analyze test questions and problems, to distribute one's time most efficiently, and to work through typical test problems. Sometimes it involves lengthy instruction indistinguishable from that of the school or college. Some forms of coaching improve predictions of performance from test scores and others do not. Unacceptable coaching methods include instruction that helps the individual use characteristics of the test items or testing situation to obtain a high score on the test regardless of knowledge of its subject matter, e.g., encouraging examinees to respond to all questions when wrong answers are not penalized, to avoid multiple-choice answers that are grammatically wrong or to use other flaws or cues in the test questions (Wigdor and Garner, 1982a: 67). When coaching is effective and some groups are more likely to receive it than others, group differences in test scores could result.

III. Sources Arising from the Test Environment

A. Effects of the Examiner

Race or sex of examiner. People may perform better on tests when the examiner is of the same race or sex as the test taker.

Language and dialect of examiner. The discrepancy between test takers' dialect or language and that of the examiner, regardless of the examiner's race, may affect test scores. Thus, the test taker could do poorly because he/she is unable to understand the examiner's oral directions or the standard English of verbal test items.

Expectancy of examiner (also known as the self-fulfilling prophecy) is the claim that teachers or examiners hold lower expectations for the performance of minority pupils than for majority pupils, communicate this prior expectation to them, and affect their test performance, resulting in the expected lower test scores.

Subjective scoring may be biased when tests are individually administered and responses are subjectively judged using scoring criteria and the test manual's examples of right and wrong answers. Bias is the scorer's tendency, when in doubt, to consistently overrate (or underrate) a given test taker's responses. If test scorers hold different expectations for various groups, they may overrate (or underrate) the responses according to those expectations. Scoring is unlikely to be a source of bias with objectively scored tests, such as machine scored, multiple choice tests.

B. Situational and Procedural Conditions

Personal tempo—a general attitude or preference for speed in doing any task that, especially on timed tests, affects test performance more than in other environments. For example, persons who answer easy items more slowly will have less time to work on harder items. Bias could occur when groups differ on this attribute.

Table 1 (continued)

IV. Sources Arising from Procedures of Test Construction and Test Use

The wrong criterion—the suggestion that group differences occur because tests are validated against an improper or irrelevant measure(s) of performance.

Overinterpretation is using test scores for applications for which they have not been validated where the group differences they show are no longer relevant. It can involve using either an unvalidated test or a validated test for another purpose. Bias as overinterpretation implies that the inappropriate application of test results produces irrelevant group differences (although the differences may have been relevant for the test's intended use).

Selection model—the use of a decision rule for selection that is perceived to place either too much or too little emphasis on test scores compared with other criteria. That is, in the context of the full range of information needed for making a decision, the information provided by the test may be given too large or too small a weight. Bias may result when the test does not show the same group differences as other valid information.

Improper standardization—the assertion that tests developed and scored using the responses of the one subpopulation (e.g., whites) are biased against another subpopulation (e.g., blacks).

Sources: These hypothesized sources were collected from Flaugher (1978), Jensen (1980), and Snyderman and Rothman (1988).

Employee Selection Procedures,[9] for example, do not include face validity among the acceptable types of validity studies.

Content validity is when a test accurately represents the relevant content domain and excludes content outside that domain. In the past, content validation often involved simply judging test items as within or outside of the content domain. More recently, it has sometimes included efforts to balance various types of content according to their occurrence within the content domain. For achievement tests, content validity may be shown by a comparison of the test content with the course of study, instructional material, and statement of instructional goals.

Criterion validity or predictive validity is when statistical analysis shows a systematic relationship[10] between test scores and one or more outcome criteria (e.g., in employment selection, elements of job performance or work behaviors). When this relationship exists, test scores can be used to predict performance.

---

[9] 29 C.F.R. § 1607 *et seq.* The Uniform Guidelines were developed by several Federal agencies and issued as regulations (see appendix A). They have been cited in numerous court cases and are required to be judicially noticed. 44 U.S.C. § 1507 (1988).

[10] Predictive validity is established using regression analysis. A "regression line" is a prediction equation of the form $Y = a + b(X)$, where "a" and "b" represent the intercept and the slope, respectively. "Y" is the criterion—some measure of academic or job performance—that can be predicted by the test scores ("X").
Differential prediction—the widely accepted definition of bias—is when the criterion scores predicted from the common regression line produce consistent nonzero errors for members of the subgroup. In this definition, the term "common" specifies that the prediction equation has been developed from the test population as a whole, proportionately representing both majority and minority groups; and "nonzero errors" refers to over and underpredictions of performance.

Although unquantifiable judgments enter into face and content validity, the statistical analysis used in criterion validation produces numbers that sharpen debates over "How much predictive validity is enough?" The degree of relationship between test score and performance is expressed as a correlation—a number ranging from 0 (for no relationship) to +1.00 (or -1.00) for a perfect relationship. Predictions are never perfect—they are often thought weak—which leads some to doubt the value and fairness of tests.[11]

Construct validity is a process of formally specifying the meaning of the measured attribute or quality. It uses a series of statements and inferences to relate what the test measures to other facts and phenomena (Cherryholmes, 1988). The question construct validation addresses is, "Do test scores relate to the world in the way they ought?"

The term "construct validity" often becomes confusing in psychological literature because the meaning and relationships of a construct can be established at many levels. Construct validation might demonstrate, at one level, that a test of problem solving ability relates to success in life, educational achievement, job performance, and social status. This is the broad meaning of construct validity that relates test scores to relevant general concepts through hypotheses and empirical evidence.

In recent years construct validation has become associated with a narrower definition incorporating the notion of job relatedness in the employment area. In this sense, construct validation entails doing a *job analysis*—carrying out an explicit set of rules to elicit descriptions of job tasks from incumbents and to ensure that employment tests for that job include measures of the tasks. The methodology of job analyses includes inferential rules for what the test should measure—for example, job tasks are what incumbents say they do on the job. However, once the job tasks are elicited, the validation process reduces to "content" validation, i.e., guaranteeing that measures of performance on the tasks that the incumbents identified make up the test. Thus, construct validation ensures that a test used, for example, to hire secretaries shows expected relationships with typing, filing, and phone answering skills.

Construct validation encompasses all other forms of validation. Content validity, predictive validity, and even face validity provide evidence in support of construct validity.

The inferential rules for doing a job analysis are well defined. Using the broader definition, what linkages between performance and test content are required to establish construct validity is ill defined. For example, what inferential rules and empirical evidence are sufficient for validating the use of a test of intelligence for broad job categories?

Minimum requirements for validation. A test is "valid" when it meets criteria established for a specific validation procedure. How demanding these criteria are, or how many different types of validity are applied, is often judgmental. Psychologists would agree that face validity alone is not sufficient. Some would argue that predictive validity is. Still others suggest that predictive validity is not sufficient without construct validity, by which they mean a job analysis.

The minimum requirements for validity seem more and more to include job analysis. The Uniform Guidelines require job analysis for both content and construct validation. However, some fear that the current emphasis on job analysis will exclude requirements for predictive validity and that job analysis may help to ensure but will not guarantee predictive validity (e.g., Gottfredson, 1988). Courts appear to be moving in this direction by asking not just for predictive validity but for construct validity too. However, many remain confused about whether requirements for construct validity mean job analysis or something more.

---

[11] Because predictions from test scores are sometimes perceived to be weak, some have suggested using social values to compensate for the low average scores of traditionally disadvantaged groups. The National Academy of Sciences' recommendation to the Department of Labor on the use of the Generalized Aptitude Test Battery (GATB) for employment referrals is one such model (Hartigan and Wigdor, 1989). Such a remedy need not presume that tests are biased, for bias is demonstrated by differential prediction, that is, *different* relationships between test scores and performance for the groups, rather than a weak relationship. These models are perceived as trading off equity against efficiency. (See the National Academy of Sciences' Interim Report on the extent of the tradeoff under a variety of models, Wigdor and Hartigan, 1988.)

Sources of Test Bias

Researchers have identified many potential sources of test bias. They look to see if test scores differ when these suspected causes are present or absent. This section discusses the many hypothesized sources of bias.

Attributes of the test takers, the testing environment, the test itself, and the statistical procedures applied during test development have all been suggested as sources of test score differences between groups.

Research studies that look for specific sources of bias may examine biases inherent in the test, such as cultural bias, or biases induced by situations having nothing to do with the test itself. Situations that can bias test results might involve, for example, the test environment, the examiner, and scoring procedures.

Hypothesized Sources of Bias. Table 1 lists a wide variety of potential sources of bias. Most of these sources merit consideration in developing, administering, or applying tests. However, one often cited as a source of bias—improper standardization—reflects a misunderstanding about test development that was already discussed and is not a proper source of bias (Snyderman and Rothman, 1988).

The factors listed in table 1 may bias test results if they affect scores in unintended ways. However, some of these factors may result in test score differences, but legitimately reflect what the test measures. For example, the following comment about the effect of test anxiety illustrates a reluctance to accept such factors as evidence of test bias simply because they affect test scores:

> [T]est anxiety is a reflection of . . . evaluation anxiety. Such anxiety can also interfere with school performance or performance on the job. Thus, anxiety effects on tests do not necessarily reduce the relationship between test scores and the outcomes observed on certain criterion measures. Indeed, the common effects of evaluation anxiety may actually enhance the predictive value of tests. . . . (Wigdor and Garner, 1982a: 67).

Other factors, such as coaching, have more ambiguous effects. Some types of coaching may produce test score differences that are biased. For example:

> coaching effects that increase test scores but not the abilities [or performance] they are intended to measure . . . affect the validity of a test.

Other forms of coaching may produce test score differences that reflect social inequalities in opportunities rather than the test's inability to predict performance.

> To the extent that coaching improves the abilities being tested and thereby improves not only the test scores but also other indicators of those abilities, then coaching is the cause of no special concern [with regard to test bias]. . . . Of course, the differential availability of coaching opportunities . . . would remain a concern. . . . But [this] concern is not fundamentally different from ones regarding other differences in opportunities such as access to private preparatory schools, to tutors. . . . (Wigdor and Garner, 1982a: 68).

In this latter example, the bias is in the availability of coaching, not in the test.

Research results on sources of bias. Research has linked some of the potential sources of bias with test score differences between racial/ethnic and gender groups. Typically these effects are small and occur with specific test content or within subgroups, rather than entire racial/ethnic or gender groups. Findings include:

- Biases due to language of the test and the examiner appear for non-English-speaking or bilingual groups. They are not found for those taught for some time in English-speaking schools.
- When items contain English words, Hispanics do better on true cognates of Spanish words and worse on false cognates. Also, Hispanics find some item types (analogies and antonyms) more difficult than other minorities (Pennock-Roman, 1991).

- Boys outscore girls when the item content is scientific, mechanical, business, practical affairs, or mathematical. Girls do better than boys when the content is human relations or the arts and humanities (Dwyer, 1976).

- Content of the Scholastic Aptitude Test (SAT) has been balanced with equal numbers of scientific, practical affairs, human relations, and aesthetic-humanities items since the 1950s (Dwyer, 1976). The apparent sex bias of the SAT suggested by its underprediction of females' college grades may occur because men and women enroll in college courses with different grading standards and levels of difficulty. Women tend to enroll in courses with more lenient grading standards. Adjusting for the strictness of grading standards reduced the SAT's underprediction of women's college performance (Strenta and Elliott, 1987; Elliott and Strenta 1988).[12]

- Practice effects are temporary and occur primarily for those who have not been tested before or recently. Recent immigrants and persons who have had little or no formal schooling or who have gone to quite atypical schools may benefit from practice in taking tests.

- Coaching for the SAT "can often produce detectable differences in students' scores especially if the students wish to improve and the instruction is good." The typical 10-point gain may help students seeking admission to highly selective colleges (Cole, 1982).

Mechanisms for Reducing Bias. Independently of research efforts to identify sources of bias, test developers have developed several mechanisms for reducing bias from the perceived sources. They have incorporated many of them in routine test development procedures. Table 2 lists many such mechanisms and the sources of bias they aim to minimize.

Many of the procedures to eliminate bias may be applied during test construction. They include reviewing test items for insensitivity, bias detection techniques, balancing items with known biases against others with opposite biases, and developing culture-reduced tests.

Sensitivity analysis is a procedure in which individuals with a variety of racial, ethnic, and gender views examine test items for insensitive language or content. Any items not meeting with approval would be eliminated or rewritten prior to test administration.

In constructing tests, test developers should carefully define test content. When specific content is not critical to the domain of the test and is known to affect test scores of some subgroups, the test may include equal numbers of test questions favoring each group. This method is known as balancing the content. If the content is critical and makes balancing it impossible (as in avoiding questions about war—a topic that may interest males more than females—on a history exam), the items may be used proportionally to their occurrence in the appropriate domain of knowledge.

Tests are sometimes subjectively rated by how specific to a particular culture their items are. Those testing information that is specific to a particular culture are "culture loaded." By contrasting culture-loaded items with items requiring only universal concepts or knowledge, researchers are exploring some ways of removing culture loading from tests. Such tests are called culture reduced. Culture-reduced tests typically present their items using pictures or symbols to avoid using language or factual knowledge that may be a product of the culture. The only prior information these tests require is an understanding of test instructions. Because the informational content of their items is general rather than specific, culture-reduced tests are best for measuring abstract reasoning or problem solving abilities rather than scholastic achievement. These tests may be particularly appropriate for use in measuring the abilities of those with a language barrier.

---

[12] Cole and Loewen debate whether the average male-female differences in SAT scores are due to bias. See the condensed transcript (part II) and their papers (part III) in this volume.

Table 2
Professionals and Test Developers' Response(s) to Potential Sources of Bias

Procedure for Minimizing Bias	*Potential Source of Bias*
I. Test Construction Procedures	
Sensitivity Analysis	Offensive language, racist or sexist language, stereotypes, and cultural bias
Bias Detection Techniques (both item bias and predictive validity)	General bias (i.e., without a specified source)
Balancing Content	Content and sex bias
Culture–Reduced Tests	Cultural bias and language bias
II. Instructions to Test Takers and Test Administrators	
Instructions Prior to Test or Test Preparation	Coaching, practice, test anxiety, achievement motivation
Instructions During Test Administration	Test anxiety, achievement motivation, reflection–impulsivity
Training and Instruction for Test Administrators	Examiner effects (e.g., expectancy and scoring effects)

Table (continued)

Procedure for Minimizing Bias	*Potential Source of Bias*
III. Professional Standards[1] and Enforcement	
APA "Guidelines for Nonsexist Language . . . "	Sexist language and stereotyping
Standards for Test Development	All sources of bias—racial/ethnic, sex, and cultural bias, differential validity, test anxiety, practice and coaching effects, personal tempo, improper standardization, the wrong criterion, etc.
Standards for Test Use	Overemphasis on test scores, overinterpretation and other inappropriate uses
Test Developer Monitoring Systems	All biases that may occur due to noncompliance with standards

---

[1] See appendix A for descriptions of standards that professional associations, government agencies, test developers, and test users have developed on test construction and use.

Test instructions, both to test takers and to test administrators, attempt to overcome many potential sources of bias. Instructions to test takers before the exam may help create the optimum amounts of test anxiety and achievement motivation. Test preparation may provide practice test questions and, consequently, experience with the test format. Instructions at the time of the test may also optimize test anxiety and achievement motivation. These instructions may encourage test takers to avoid impulsive answers and read all responses before choosing the most correct one (reflection-impulsivity).

Instructions to test administrators encourage them to follow proper procedures in administering tests, including having a neutral attitude toward test takers.

Professional standards reinforce the use of many of these mechanisms for reducing biases. For example, the American Psychological Association (APA) Standards (described in appendix A) reduce potential biases of the test administrator or testing environment by encouraging test use only by trained professionals (APA, 1985, Standard 6.6). Test administrators should provide an environment free of distraction (Standard 15.2) and follow proper procedures regarding instructions to test takers, time limits, and scoring (Standard 15.1).

APA standards also discourage biases from practice and coaching (Standard 3.14, "The sensitivity of test performance to improvement with practice, coaching, or brief instruction should be studied. . . .); personal tempo (Standard 3.13, "For tests that impose strict time limits, test development research should examine the degree to which scores include a speed component. . . . "); and race, sex, and cultural background (Standard 3.5, "When selecting the type and content of items . . . test developers should consider . . . cultural backgrounds and prior experiences of the variety of ethnic, cultural, age, and gender groups. . . .).

The APA first published its "Guidelines for Nonsexist Language in APA Journals" (1986) in 1977. They describe ways in which writers can avoid sexist language or stereotypes generally. The guidelines apply equally well to the writing of test items.

The Joint Committee on Testing Practices[13] recently developed a "Code of Fair Testing Practices in Education." It addresses many similar issues. Concerning overinterpretation, test developers should "[w]arn users to avoid specific, reasonably anticipated misuses of test scores"; and test users should "[a]void using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use" (B–11).

Professional associations are not, however, the only groups establishing standards. The Federal Government and test developers and test users have also written standards for test construction and use (see appendix A). All of them have based their guidelines on the APA standards.

Despite the many sets of standards, few mechanisms are in place to monitor compliance with the standards. The EEOC's Uniform Guidelines on Employee Selection Procedures instruct employers on the proper legal use of tests and other selection procedures. Legal and regulatory channels enforce them. Other standards, however, have no obvious means of enforcement. The Educational Testing Service, which develops the SAT and many occupational licensing exams, has established its own (both internal and external) monitoring system to ensure compliance with the standards, but such practices may not be common among test developers.

Appropriate Use of Tests

The issue of whether tests are valid is addressed by assessing the evidence that they have been properly standardized, examined for and purged of bias, and justified with predictive, content, and/or construct validity. But the conclusions about validity refer only to the application(s) included in the validation studies. Rather than ask, "Is the test valid?", a more appropriate question is, "Valid for what?", because a test can be validated for one purpose and used for another. Here "inappropriate use" distinguishes abuses of testing in its applications from inadequate test development procedures.

---

[13] The Joint Committee is a cooperation of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

Inappropriate uses of tests include overinterpretation or using validated tests for purposes other than those for which they are validated; using a biased test instead of an equally valid, but less biased one; and superfluous use of tests.

Many of the attacks on testing today concern overinterpretation. In education, some examples of inappropriate uses are using student achievement tests to evaluate teacher performance or school effectiveness; using aptitude tests, such as the SAT, to award scholarships based on achievement; and using tests developed to select students into teacher education programs to hire, promote, or set salaries for teachers. Any of these uses would be appropriate if a study were done or evidence documented showing that the test is valid for the expanded use. In the employment area, what constitutes overinterpretation is debated (see Schmidt, 1988). First, are tests situation specific, that is, valid for a job in one organization or setting and invalid *for the same job* in another organization or setting? Second, are tests of cognitive abilities job specific, that is, valid for some jobs and invalid for others? "Validity generalization" is the attempt to generalize from validity studies of tests of cognitive abilities performed on a representative sample of jobs to all jobs, even those at the lowest skill levels. Where employers can use other studies to validate tests without overinterpreting test results is unclear.

Among tests that are otherwise equal, using a test that has greater adverse impact is not only inappropriate but against Federal regulations. The Uniform Guidelines[14] state: "Where two or more selection procedures are available which serve the user's legitimate interest in efficient and trustworthy workmanship, and which are substantially equally valid for a given purpose, the user should use the procedure which has been demonstrated to have the lesser adverse impact."

There are also meaningless ways of using tests, such as testing for testing's sake or using tests to sort or label people without applying the appropriate intervention. The Department of Education has challenged the use of tests, not because the test was biased, or even that it disproportionately assigned minority students to special classes, but because the classes those students were assigned to could not be justified as advancing their education. They were not designed to fulfill the potential of the students in them and were seen as a dead end.[15] Such testing is also inappropriate.

Methods of Overcoming Perceived Bias or Adverse Impact

Many ways have been suggested for overcoming test score differences and the resulting adverse impact. They arise from perspectives reflecting different definitions of bias and different levels of confidence about the existing information on test bias. These perspectives include 1) those who think tests are biased and that bias and adverse impact should be eliminated; 2) those who are uncertain about whether or not tests are biased, but believe adverse impact must be eliminated; 3) those who think that tests are unbiased, but adverse impact should be eased or eliminated, nonetheless; and 4) those who conclude that tests are unbiased and that eliminating the adverse impact of tests (apart from improving the validity of tests) is inappropriate. The choice of solution depends upon these views. In particular, are there solutions that are appropriate when tests are biased and inappropriate if the tests are valid?

---

[14] 29 C.F.R. § 1607.3.

[15] See the U.S. Department of Education's Administrative Proceeding against Dillon County School District, South Carolina (Docket number 84-VI-16). The use of the test to assign students to dead-end classes may also have been a concern in *Larry P.*

The solutions reviewed below are quite varied. Some suggest changes in a test or its scoring; others propose a different emphasis on skills and abilities. Each solution must be evaluated in light of assumptions about the validity of tests, practicality, and the unintended consequences and costs to test users and society.

Proscribing the use of tests. When Judge Peckham *(Larry P.)* perceived an IQ test as biased, he banned its use for placing black children in classes for the mentally retarded. When the ban was later expanded, unintended effects were reported. "Now, no black children may be given IQ tests for any purpose, nor may they have the results of privately or out-of-state administered tests entered in their school folders . . ." (Elliott 1988). Thus, the ban restricted use of blacks' IQ test scores for evaluation, admission, and placement in special classes for the retarded or learning disabled. "School psychologists now use bits and pieces of various tests without benefit either of norms or of validation." Finally, students "too slow for the mainstream classes but not fitting the requirements for service as learning disabled children . . . are left to flounder, and sometimes founder, in regular classes." In the meantime, black overrepresentation in classes for the educable mentally retarded and learning disabled continues (Elliott 1988).

Furthermore, in *Crawford v. Honig,*[16] the mother of a black child in a special education class sought to have her child tested because she believed the child could be moved into a normal curriculum. The Federal court's order in *Larry P.*, however, prevented her from having the child tested, with the net result contrary to what that decision intended. *Crawford* challenged the expansion of the ban on testing to additional purposes and to the learning disabled. Judge Peckham lifted the ban against using tests for placing black children in mentally retarded classes in September 1992.

Alternative criteria. Alternative criteria showing less adverse impact than tests have often been proposed. Table 3 lists many criteria for employment selection, some of which include tests as part of the assessment (from Reilly and Warech, 1991). The utility of other criteria may depend upon whether or not they predict performance as well as the tests they replace. Switching from a test to alternatives that are not as valid as tests may lower selection performance standards and increase the costs of productivity and errors for the employer (Gottfredson, 1988). One obvious solution to this dilemma is to find "selection systems with *reduced* adverse impact and *enhanced* utility" (Schmidt, 1988).[17]

In reviewing evidence for validity, adverse impact, and fairness of each alternative, Reilly and Warech conclude that trainability tests, work samples, biographical history (called "biodata"), and assessment centers may have the desired properties—greater validity and less adverse impact than cognitive ability tests. Personality tests, self-assessments, training and experience evaluations, expert judgment, seniority, handwriting analysis, and reference checks were clearly inferior in validity to ability tests. (Tests of honesty and physical ability were appropriate for specific types of jobs.)

Are trainability tests, work samples, biodata, and assessment centers practical and viable alternatives or supplements to mental tests and will they aid in reducing adverse impact?

Emphasizing multiple skills, attributes, and abilities throughout society. Other researchers believe society should emphasize multiple skills, attributes, and abilities in education, job selection, and other societal rewards. Proponents of this solution suggest that the concept of performance itself and society's economic reward system is too narrowly based upon a single ability, IQ. They encourage the use of a combination of tests and the alternatives listed in table 3 because the alternatives represent other skills (e.g., interpersonal ones) that are not measured by

---

[16] Crawford v. Honig (C-89-0014 RFP). Also see, "Judge Lets California Resume IQ Testing of Black Students," *Education Daily*, Sept. 8, 1992, p. 4.; and "Judge lifts ban on IQ testing," *The Washington Times*, Sept. 3, 1992.

[17] Another solution is to establish policy goals that are not strictly related to short-term productivity or efficiency.

Table 3
Alternatives to Cognitive Ability Tests in Employment

Work samples—having the applicant do a task or set of tasks that, based upon systematic job analysis, is directly relevant to the job.

Assessment centers—a comprehensive, standardized procedure using multiple assessment techniques in combination, including situational exercises and job-related simulations as well as paper and pencil tests and interviews. These are expensive and are mainly used for staffing managerial positions.

Trainability tests (also known as minicourses and miniaturized training and evaluation tests)—the presentation of job-related training materials followed by an assessment of learning with a paper and pencil or performance test.

Evaluation of past training and experience

Reference checks—information obtained from a previous supervisor regarding the performance of the applicant under job relevant conditions.

Seniority

Grades and educational achievement

Individual assessments by experts—the use of expert judgment (e.g., a consulting psychologist) to combine and summarize objective data.

Interviews—an oral interaction (either unstructured or highly structured, resembling an oral test) between a job applicant and a representative of the employer used to predict job-related behavior.

Peer evaluations—judgments about a job candidate by co-workers or co-trainees.

Biodata—biographical information typically collected as part of a standard application form and formatted to allow objective classification of responses. Information may be objective and verifiable or difficult or impossible to verify.

Self-assessments—an individual's self-evaluation of ability, skill, knowledge, or other traits.

Personality tests

Projective techniques—processes that measure personality dimensions in a disguised fashion by presenting ambiguous stimuli to which examinees respond in an unrestricted format.

Handwriting analysis

Tests of physical ability

Honesty tests—either the polygraph or paper and pencil tests designed to identify dishonest job applicants.

Source: See the review of alternatives in Reilly and Warech (1991).

aptitude and achievement tests. They also support efforts to identify and use other skills and abilities that show smaller differences between groups.

The Golden Rule Procedure. This controversial test construction procedure[18] (discussed more fully below) eliminates test questions that show the largest differences between groups. Whether it is appropriate may depend on whether tests are biased or valid. If tests are truly biased, eliminating test items showing major group differences would seem to reduce group differences. If tests are not biased, however, eliminating such items may not be an appropriate or effective way of handling adverse impact. This procedure may unintentionally diminish the predictive validity and utility of tests. Some suggest it may result in tests constructed of easy items, identifying test takers with minimum competency rather than those able to handle tougher everyday tasks or rare but critical situations. Uncertainty about the extent of bias in tests suggests that more information is needed about how much this procedure reduces adverse impact and changes the predictive validity and value of the test.

Using Minimum Competency Standards. Selecting those who meet minimum competency standards rather those who score highest on tests is a way to increase the number of eligible individuals from lower scoring groups. Moreover, when a test measures only part of the relevant domain of skills, knowledge, or abilities, minimum competency standards may be appropriate for identifying a pool of candidates for further evaluation.

The effects of minimum competency standards, however, are hotly debated. According to Schmidt (1988: 288), they reduce selection and performance standards for all applicants, majority group members and minority group members, and lead to large losses in productivity across the entire work force. Some, however, dismiss Schmidt's evidence as weak.

"Within-group scoring" is assigning test scores relative to each person's gender, race, or ethnic group. When applied specifically to races, it is known as race norming. Within-group scoring effectively selects the highest scoring minorities in proportion to their presence in the applicant pool or reference group.[19] The Civil Rights Act of 1991 recently outlawed the use of any such adjustment (i.e., on the basis of race, color, religion, sex, or national origin) for the selection or referral of applicants or candidates for employment or promotion.[20]

Selection rules incorporating values. Selection rules intended to overcome adverse impact combine test scores with social values. These rules recognize that tests, even if unbiased (i.e., the test scores predict the same performance regardless of race), do not predict performance perfectly. Social values are allowed to influence selections to some extent because test score predictions are less than perfect, that is, some individuals who could perform well will have low test scores.

Using minimum competency standards or within-group scoring or incorporating social values into the selection rules are methods perceived as trading off the efficiency of the work force for equity. Some of these methods, however, have larger effects than others. For example, the adjustments to test scores that are applied when social values are taken into account are typically much smaller than those applied by within-group scoring. But again, the use of such adjustments on the basis of race, color, religion, sex, or national origin is restricted by the Civil Rights Act of 1991.

---

[18] The procedure is named for the law suit that stipulated it as a condition of settlement: Golden Rule Ins. Co. v. Washburn, 1984 (No. 419-76, Ill, 7th Jud. Cir.).

[19] Some have pointed out unintended negative consequences of certain affirmative action programs that they claim lower selection requirements for minority groups, like race norming. For example, Steele (1989) suggests they may reinforce the myth of black inferiority, particularly on campuses. Blacks may often enter college with lower test scores and high school grade point averages and with less college preparation and poorer schooling in relation to their white counterparts. They generally get lower grades, fail, and drop out at higher rates than the better prepared whites. The better prepared blacks who can compete are often perceived as affirmative action cases. If, as Steele suggests, these programs raise the possibility of perpetuating a myth of inferiority and forcing more qualified minority members to endure stereotyping, such effects must be carefully weighed against the college benefits to students who would not have been in college without the affirmative action program.

[20] Pub. L. No. 102-166 (Nov. 21, 1991) Sec. 106.

Race-Neutral Meritocracy. A race-neutral meritocracy would seek to apply valid tests in appropriate ways without regard to race, ethnic group, or gender and despite the effects of adverse impact.

Issues

Internal Validation Issues. Internal validation examines the properties of the tests themselves to identify test questions that represent extraneous factors, such as bias. These procedures look at how different demographic groups perform on the test items.

> *Issue 1: How should test items that are biased be identified? Is it sufficient that an item is more difficult for one group than another, or should item difficulties only be compared for test takers with the same test score?*

Most testing experts believe that group differences in correct response rates are evidence of bias only if the groups have the same average ability. Thus, the generally accepted method of identifying biased test items assumes that systematic biases have first been removed from the test and then examines group differences in the difficulty of test items among subgroups of test takers having the same total test score. Despite the wide acceptance of this definition, differences in how difficult the items are for various groups are sometimes regarded as evidence of bias. In particular, lawsuits have relied upon definitions of bias that did not share widespread acceptance.

One suit[21] challenged an Illinois insurance licensing exam for its validity and intentional discrimination. The out-of-court settlement assumed that differences between groups in item difficulties were bias. It specified the test construction methods ETS must use when developing the licensing exam.[22] The method is known as the Golden Rule procedure. It dictates the order in which items selected from an item pool can be included in the test and thereby minimizes the number of test items that are significantly more difficult for minority groups than for whites. In selecting items with moderate to high difficulties, ETS must first use those with the smallest black-white differences.[23] It must justify passing over any such items.

Critics of the Golden Rule procedure have identified several problems with it. Suppose that groups actually differ in their potential performance. Then item difficulties must be able to vary across groups. Applying the Golden Rule procedure, however, would eliminate items that allow for differences between groups in average ability. It could, therefore, produce less valid tests because the items that best measure underlying differences in performance had been eliminated.

Gottfredson (1988) suggests that tests constructed using the Golden Rule procedure would be very easy ones. In the employment area, they would likely be composed of items representing frequent, routine job tasks rather than critical ones where errors are quite costly to employers. Although tests constructed using this procedure could be content validated through a job analysis, she says, predictive validity and test utility may be impaired.

---

[21] Golden Rule Ins. Co. v. Washburn, 1984 (No. 419-76, Ill., 7th Jud. Cir.). See also Allen v. Alabama State Bd. of Ed., 612 F. Supp. 1046 (M.D. Ala. 1985), *vacated*, 636 F. Supp. 64 (M.D. Ala. 1985), *rev'd*, 816 F.2d 575 (11th Cir. 1987). The latter case's settlement also specified the method for assembling test questions to form tests. The required method is a variant on the "Golden Rule procedure" that was specified by the former settlement.

[22] FairTest has been working to extend the application of the Golden Rule Procedure ("FairTest Examiner," 1987: 4).

[23] Items with group differences in response rates that are less than 15 percent must be used first.

Anrig (1987) points out that if followed exactly, the Golden Rule procedure violates many common sense rules about test construction. An item that reveals the answer to another would be included on the same test if it met the Golden Rule criteria. Also, similar items could not be reserved for parallel forms of later tests but must appear in the same test.

Except for the test involved in the settlement, ETS compares the difficulty of items among blacks and whites (and other subgroups) who have the same test scores.

Other legal decisions have also been behind the state of the art in the methodology they used. In *PASE v. Hannon*,[24] Judge Grady "thought that since tests are made up of items, it is in the items that bias will or will not be found. He begged each side to supply him with item analyses. Neither did . . . . " (Elliott, 1988: 338). Grady then examined the several hundred items on three tests, and identified nine items that were biased according to his personal judgment. This technique, known as "face validity," is generally perceived as inadequate because individuals seldom agree on which items are unfair. This case did not advance understanding of which item statistics provide acceptable evidence of bias.

Issue 2: Should biased items be categorically eliminated from tests or kept in when they are strongly related to what the test measures or balanced with items having an opposite bias?

An item statistic calculated for internal validity studies is a *point-biserial correlation*. In this context, it is a special correlation showing the relationship between an individual's response to an item and his total test score. A "large" correlation indicates that the item is a good measure of the phenomenon represented in the test as a whole, that is, good at distinguishing high scorers from low scorers.

When items have a strong relationship with the test content, test developers often keep them in the test despite large minority or gender group differences in item difficulties. Omitting these items would limit the test's ability to measure performance and could result in a very poor test. For example, males tend to do better than females on test items about war, conflict, or aggression. A history test with items on war would show gender differences in item difficulty for such items. However, could an achievement test in history adequately reflect the content of the subject without including some items on war? Test developers often use judgment in deciding when content showing group differences is inherent to the test.

When biased items cannot be eliminated, they can sometimes be balanced with the same number of items with the opposite bias. This solution, however, raises questions about the test content. If girls do better on questions about the arts and humanities and boys do better on questions about science and technology, are such questions most appropriately counterbalanced or justified according to their frequency in some larger domain of knowledge or information? (See issue 8, below.)

Issue 3: What proportion of test items in current tests is biased (using any of the above definitions)?

The proportion of items that is biased depends upon the bias detection method and its underlying definition of bias. In a typical test, 70 percent of test items may be identified as biased using the Golden Rule procedure. On the other hand, methods that first use overall test score to control on ability may flag many fewer test items as questionable. Furthermore, some experts think that the items identified by these methods should be scrutinized, but are not necessarily biased. Recent changes in terminology reflect this viewpoint by referring to "differential item functioning" instead of "item bias" to describe test questions that may not be biased but may have different meanings for various groups.

---

[24] 506 F. Supp. 831 (D.C. Ill. 1980).

Issue 4: How much does eliminating items identified as biased reduce test score differences between groups?

Seldom do allegations that tests are biased quantify the extent of that bias. Sometimes the extent of bias is characterized by identifying a number of test items that are thought biased. However, removing the items that are thought to be biased has not always reduced the average test score differences between groups.

For example, Anrig (1987) compared the results of tests assembled according to "traditional" procedures and those assembled according to the Golden Rule procedure. The Golden Rule procedure had the potential of increasing the proportion of blacks who passed the Illinois licensing exam because it put items with the largest race differences in correct-answer rates at the bottom of the test developer's pile where they were unlikely to be included on the test. However, Anrig claimed that tests developed using the Golden Rule procedure did not affect the passing rate among blacks. Other methods of identifying biased items may be more effective in reducing differences between groups.

External Validation Issues. External validation requires, first, establishing predictive validity—that test scores do predict performance—and, second, showing that the predictions are the same regardless of group membership. Differential prediction (as subgroup differences in predictive validity are known) is a widely endorsed definition of bias.

Issue 5: Is the predictive validity of tests the same for different racial/ethnic and gender groups?

Despite early concerns that tests may not predict the same performance for some groups as for others, testing research shows that in most instances differential predictive validity does not exist. "Across numerous studies and contexts the [statistical relationship between test scores and performance (i.e., predictive validities) for blacks and whites] either do not differ significantly or the bias is in favor of blacks" (Shepard, 1987).

> Whether the criterion to be predicted is freshman GPA in college, first year grades in law school, outcomes of job training, or job performance measures, carefully chosen ability tests have *not* been found to underpredict the actual performance of minority group persons. . . . [T]he bulk of the evidence shows either that there are essentially no differences in predictions based on minority or majority group data, or that the predictions based on majority group data give some advantage to minority group members. In most instances, the use of separate equations for . . . selection would reduce, rather than increase, the number of minority group members selected (Linn, 1982: 384–85).

Issue 6: How high should correlations of test scores with performance be for a test to be "valid"?

Predictive validity is measured with a correlation, an index that measures the degree of relationship between test score and performance on a scale of 0 (no relationship) to +1.00 (or −1.00). A high correlation between test scores and performance is desirable because it suggests that the test score predicts performance very well. Because test scores' predictions of performance are less than perfect, many question the value of tests and whether tests are useful enough to outweigh the effects of adverse impact on minorities. Others suggest that less than perfect predictions could result from poor measures of performance rather than the problems with the tests. Thus, the size and meaning of the correlations between test scores and performance are frequently debated.

Seymour (1988) argues that correlations of 0.2 and 0.3 are too small to be of significance. Schmidt (1988: 278–79) argues that even when the validity of a test is low, it "is still large enough to be of practical value in selection. (A validity of .23 has 23 percent as much value as perfect validity, other things equal.)" Gordon, Lewis and Quigley, in a rebuttal to Seymour, suggest that in a highly competitive world, the edge provided by even a weak measure of productivity may be critical to the survival of a business.

Schmidt and Hunter (Schmidt, 1988; Schmidt and Hunter, 1981) have developed a method for estimating the payoff to employers of increases in employee job performance. Schmidt concludes that "selecting high performers is more important for organizational productivity than had been thought." "[F]ailure to use [cognitive employment tests] in selection will typically result in substantial economic loss to individual organizations and the economy as a whole" (1988: 280–81). However, compared to typical economic analyses of productivity, their methods are very crude and many believe that more analysis needs to be done on this question.

Discussions such as these often assume that the test fails to predict performance because it is flawed. Alternatively, the measure of performance itself may be flawed. For example, the measure of performance may be based upon subjective judgements or it may represent too many or too few of the relevant skills and knowledges. In such instances, an imperfect measure of performance could result in a low correlation with test scores, even if the test is very good at predicting the actual (rather than measured) performance.

> *Issue 7: If predictive validity of a test is high and the same across groups, is it also necessary to establish other types of validity (e.g., content validity or job relatedness)?*

Many regard predictive validity as sufficient for validating a test. Linn (1980: 522), however, describes a growing consensus that content, predictive, and construct validity should "be viewed as approaches to accumulating certain kinds of evidence rather than as alternative approaches, any one of which will do." Recent interpretations of the Uniform Guidelines and some litigation[25] also show an emerging trend where predictive validity is no longer sufficient without, for example, a job analysis and construct validity. That is, the statistical relationship between a test score and an overall measure of performance is no longer adequate without a job analysis or other means of linking test material to job duties or other components of performance.

Others suggest that the method of validation depends upon the type of test. "Criterion-related validation strategies are more possible for employment or college admissions tests . . . because usable criterion measures are usually more readily available. . . . [L]icensing and certification tests . . . lend themselves to . . . a content validation strategy . . . augmented by evidence of construct validity" (Madaus and Shimberg, 1989). Furthermore, if construct validity is established using a job analysis, is predictive validity necessary? The minimum requirements for test validation remain unclear—whether they are predictive validity, predictive validity and some other form, a job analysis by itself, a job analysis and predictive validity, or some other approach.

> *Issue 8: Job relatedness is usually established by doing a job analysis; content validity by doing a content analysis. If these analyses are necessary, what procedures should be followed in conducting them? For example, how should those who contribute to job analyses be selected?*

A job analysis is a procedure whereby researchers elicit job tasks from a panel of incumbents to ensure that employment tests for that job include measures of the performance of, potential to do, or knowledge of those tasks. What criteria should be used to select incumbents to participate in the job analysis?

In educational or employment testing, how should the content domain be defined and questions apportioned among topics?

Truth in Testing Issues. Several issues have emerged in the "Truth in Testing" movement. For the most part, these issues are concerned with protecting test takers, rather than validity per se. However, exceptions are described below.

The movement includes a variety of efforts to regulate standardized testing through State or Federal government. The proposed regulations would require that:

(a) individual test takers have access to corrected test results within a specified period after test administration;

---

[25] See, for example, Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975).

(b) test sponsors or publishers file information on test development, validity, reliability, and cost with government agencies; and

(c) testing agencies give individual test takers information on the nature and intended use of tests prior to testing and guarantee their right of privacy concerning their own test scores (Haney, 1981).

Protecting test takers through such regulations may create problems for test developers. In particular, part (a) requires test developers to publish their tests such that they must continuously develop new ones. Developing and validating new tests is very time-consuming and costly. The cost may be especially acute for specialized exams, such as licensing exams, where test content is fairly stable and the pool of test takers small.

If test developers felt compelled to control costs by reusing published test items, test validity would be impaired. Examinees coached with practice on published tests could score higher but would not necessarily perform better when selected according to their test results.

Another issue concerns the section of trial items that test developers often include in ongoing testing programs. Should test developers, such as ETS, be required to designate the pilot subsections of tests so that test takers may skip them (because they will not be counted in the scoring)? Those favoring this requirement argue that test takers may tire during the test and would do better if they could omit unscored sections. Test developers fear that the respondents who choose to answer optional sections will not represent the total group of test takers and will create biases in the pilot results. This problem could result in poor items appearing in later editions of tests.

Rhode Island and New York are two States that have already passed legislation addressing some of the Truth in Testing concerns. The Ford Foundation is supporting a study of the feasibility of establishing a national regulatory agency to monitor testing and protect test takers.

Legal and Policy Issues. Tests and test use have been the subjects of legislation, litigation, and Federal regulations and State control increasingly more often in recent years.

Efforts of the "Truth in Testing" movement have led to State legislative proposals to form State advisory committees that would review the effects of standardized tests on test takers of varying racial, ethnic, linguistic, and gender backgrounds and consider methods of assuring fairness and equity of such tests. These proposals specify the analyses test developers must do in examining tests for bias and then report to the committees.

Most court cases challenging tests have involved allegations of discrimination in employment decisions to hire or promote workers. What proof employers must have to demonstrate that the tests they use to select or promote employees are fair has gradually evolved with some important recent changes.

The use of tests in education, for pupil assignment in schools, has also been challenged in and out of court, the latter by the Office for Civil Rights in the Department of Education. Appendix B describes some of the more important court cases and out of court settlements.

The "Uniform Guidelines on Employment Selection Procedures"[26] have also been cited in several court cases and have had an enormous effect on employers' employment selection procedures.

These activities have resulted in several legal and policy issues.

Issue 9: What are the legal and policy issues relating to the development and use of tests?

Some especially timely legal issues are described below. They involve the debate over who has the burden of proof and evidentiary standards in cases alleging employment discrimination; the Department of Labor's pilot testing a method of scoring tests separately by race when referring job candidates to employers; and the State of California's refusal to provide or use testing services in assigning black children to classes because of the statewide ban on IQ testing. Finally, have courts concluded that tests were biased or used inappropriately?

---

[26] 29 C.F.R. 1607.

Burden of Proof and Evidentiary Standards. Tests are frequently the job selection criteria used in cases alleging employment discrimination. The 1971 case, *Griggs v. Duke Power Company,*[27] established the concept of disparate impact, whereby absent proof of discriminatory intent, an employer could still be found discriminatory based upon the consequences of employment practices (including the use of tests). The standards of proof for showing that tests are valid for selecting and promoting employees were clarified and extended in later cases and, except for a brief interlude between recent Supreme Court decisions and the passage of the Civil Rights Act of 1991, have undergone few changes since then. They require that the plaintiff must first establish a prima facie case of discrimination; then the defendant employer must demonstrate that the test is a business necessity (i.e., demonstrates a manifest relationship to the employment in question); and finally, the plaintiff may prevail by offering either an equally effective alternative practice that has a less discriminatory impact or proof that the apparently legitimate practices are a pretext for discrimination. Courts have relied upon the "Uniform Guidelines on Employment Selection Procedures" for standards on whether tests are a reasonable measure of performance.

Recently the Supreme Court extended the disparate impact analysis that applied to tests to subjective measures of job performance in *Watson v. Fort Worth Bank.*[28] Then, in *Wards Cove v. Atonio,*[29] it further developed the theory of discriminatory impact. In particular, the Court shifted the burden of proof on the business necessity issue from the employer to the plaintiff and changed the standards of evidence required of the parties.

In *Wards Cove v. Atonio*, the Court adopted standards for the evidentiary burdens applicable to employment discrimination cases that had been enunciated earlier by a plurality in *Watson v. Fort Worth Bank.* The Court agreed that statistical disparity is not sufficient to establish a prima facie case and that the plaintiff must identify the specific employment practice or practices responsible for the disparity and prove that each employment practice separately causes a disparity. After the employee(s) establish a prima facie case, the employer may refute the statistical evidence by pointing out fallacies and deficiencies or demonstrate legitimate business reasons for the employment practice. However, the practice need not be "essential" or "indispensable" to the employer's business.

Because the Supreme Court decisions were seen as substantially weakening the standards of proof and evidence established by *Griggs,* Congress passed the Civil Rights Act of 1991.[30] This Act restores the burden of proof and standards (e.g., the concepts of "business necessity" and "job related") prevailing before the *Wards Cove* decision. Also, it clarifies that the complaining party must demonstrate a disparate impact for each particular challenged employment practice, except if he demonstrates that the elements of an employer's decisionmaking process are not capable of separation for analysis, he may analyze it as one employment practice.

Because the Civil Rights Act of 1991 overrides most of the effects of *Wards Cove*, the primary recent change with respect to testing is *Watson's* extension of validation procedures to subjective measures of performance. Subjective criteria for selection, a viable alternative to tests in the past, will now require justification or validation when they show adverse impact. Employers' selection procedures may change. More generally, the passage of the Civil Rights Act of 1991 strengthens the deterrents against discrimination and provides better protection for those who suffer employment discrimination.

---

[27] 401 U.S. 424 (1971).

[28] 487 U.S. 977 (1988).

[29] 490 U.S. 642 (1989).

[30] See *Report of the United States Commission on Civil Rights on the Civil Rights Act of 1990,* (Washington, DC: U.S. Commission on Civil Rights, July 1990). This report analyzes the changes wrought by the Supreme Court decisions and the intent of the provisions of the Civil Rights Act of 1990, most of which were subsequently adopted in the Civil Rights Act of 1991.

Within-Group Scoring or Race Norming.[31] The United States Employment Service (USES), under the auspices of the Department of Labor, administers and scores the General Aptitude Test Battery (GATB). Those seeking referrals to employers take this test and have their results communicated to employers. Since 1981, however, the USES has operated pilot studies that score the tests separately within racial/ethnic groups—blacks, Hispanics, and all others—and refer the highest scoring within each race regardless of how their scores compare across races. Thus, different racial groups are scored differently with respect to each other and the same with respect to individuals within their group. This feature is called "within-group scoring" or "race norming."[32]

The United States Department of Justice challenged the practice of within-group scoring on both constitutional and statutory grounds. It charged that within-group scoring constitutes intentional racial discrimination in that it prefers some and disadvantages other individuals based on their membership in racial or ethnic groups (Delahunty, 1988).

The Department of Labor sought guidance from the National Academy of Sciences (NAS) on the consequences for economic efficiency and social equity should the method be widely adopted. NAS issued its report, *Fairness in Employment Testing* (by Hartigan and Wigdor), in June 1989. The report evaluated the validity of the GATB and made its own recommendations about how test scores should be adjusted for fairness. The study did not, however, address the legality of race norming.

Still faced with the Department of Justice's challenge, the Department of Labor proposed a moratorium on the use of the GATB for a broad range of jobs[33] as well as the allegedly unconstitutional scoring technique. The Civil Rights Act of 1991 recently outlawed the use of within-group scoring. However, because the Department of Labor has not yet implemented its moratorium with a final rule, States have no clear guidance on whether the GATB will be supported as valid should its use with a broad range of jobs be challenged for adverse impact of the unadjusted scores.

Banned Tests for Particular Groups. Larry P.[34] banned the use of IQ tests in California schools because the court viewed the test as biased against black children. The children who scored low on the test were placed in special education classes for slower students. Should the use of a test be banned for particular groups when they are judged to be biased, unfair, or unjustified educationally?

---

[31] References for the following statements include: Department of Labor, Employment and Training Administration, "Proposed Revised Policy on Use of Validity Generalization-General Aptitude Test Battery for Selection and Referral in Employment and Training Programs; Notice and Request for Comments," *Federal Register*, Tues., vol. 55, no. 142, July 24, 1990, pp. 30162–30164; Frank Swoboda and Judith Havemann, "Labor Dept. Abandoning Blue-Collar Aptitude Test," *The Washington Post*, July 11, 1990; U. S. Department of Labor, Employment and Training Administration, News Release (USDL: 90–354), "Dole Suspends Use of Job Aptitude Test;" and conversations with Department of Labor officials.

[32] The post-1981 GATB had another new feature. The number of jobs with GATB-based referrals was expanded from the 450 for which the test was originally validated to perhaps the whole 12,000 jobs in the economy. The expansion was justified using indirect evidence of validity. This feature is known as "validity generalization." The race-norming feature was added to compensate for any adverse impact occurring with the expansion.
Validity generalization has been challenged by many who doubt that the GATB is relevant for all 12,000 jobs in the economy. Unlike within-group scoring, however, validity generalization has not been challenged within the context of our legal system.

[33] The Department of Labor has continued to support the use of the GATB in its pre-1981 form, i.e., for the 450 jobs for which it was originally validated (and without within-group scoring).

[34] *Larry P. v. Riles*, 495 F.Supp 926 (N.D. Cal. 1979), *aff'd in part, and rev'd in part*, 793 F.2d 969 (9th Cir. 1984).

Bias v. Inappropriate Use. In *Larry P.*, the court struck down the use of IQ tests for black children as biased. Other courts have looked at tests and found they were inappropriately used for a purpose other than that for which they were intended. More generally, have courts concluded that tests were biased or inappropriately used?

Issue 10: What effects have legislation, litigation, and government regulations had on testing?

Some suggest that legislation, litigation, and government regulations have helped eliminate unnecessary test use and promote the use of alternatives with less adverse impact. The beneficial societal effects that have occurred include the promotion of equality among groups, a reduction in racial tensions, a realization of the productive potential of citizens who previously would have been barred from opportunities because of test results, and increased productivity of all citizens.

Others claim that more stringent standards for, or restrictions on, test use have had dire consequences. Some of the suggested consequences are:

- Employers have difficulty validating selection criteria because the "Uniform Guidelines" have established, as minimum requirements, standards that were intended as ideals.
- In States that have banned test use in schools, educators have been piecing together portions of tests and making selections without validation.
- Selection criteria that deemphasize knowledge, skills, and abilities have reduced the United States' productivity.
- Proposed State legislation that requires test developers to publish their tests typically 2 years after their use will increase test development costs substantially, especially in some smaller testing programs. Requirements to identify pilot subsections will undermine the development of unbiased tests in the future.

What evidence or solutions are there for each of these perspectives? Do the beneficial effects outweigh the more harmful ones?

Issue 11: What influence has social science had on legal and regulatory processes?

Have legislators, courts, and regulatory agencies responded to state-of-the-art social science in handling testing issues? Some instances suggest that the flow of information between social scientists and those who make law (and vice versa) is uneven. Judge Grady's inability to obtain item analysis from the defendant or the plaintiffs in *PASE* despite the common use of such techniques at the time is an example where the evidence presented in court was behind the state of social science. The EEOC's issuance of, and courts' reliance upon, the "Uniform Guidelines" as minimum requirements when psychologists regarded them more as ideals is an instance where litigation appears to be taking the lead. Are there other such disjunctions?

Have advances in social science research changed perspectives on any court decisions, legislation, or regulations?

Issue 12: What legal or regulatory changes should be made concerning testing?

Should State or Federal agencies or advisory committees be established to monitor testing and review methods used to achieve equity? Should test developers be required to report analyses of tests and test items by race, ethnicity, and gender to a public agency for review? Should legislation dictate the definition of bias that test developers use in choosing items for their tests?

Many have suggested that the EEOC's "Uniform Guidelines" should be revised because they are too demanding. However, these guidelines have been in place now for over 10 years. During that time many test users have made a concerted effort to conform to the guidelines. Considering this effort and other developments in validation procedures, are revisions appropriate? If so, how should they be revised?

Part II

Condensed Transcript of the Consultation

Consultation on the Validity of Testing in Education and Employment
U.S. Commission on Civil Rights
Friday, June 16, 1989

The following is an abbreviated version of the transcript from the consultation. The text has been condensed and reorganized, but every effort was made to preserve its meaning.

VICE CHAIRMAN FRIEDMAN: Good morning. Welcome to the consultation of the Civil Rights Commission on the Validity of Testing in Education and Employment. I want to welcome our guests and experts here today and introduce the Commission staff who are here. Melvin Jenkins is the Acting Director of the Civil Rights Commission. Kim Cunningham is in charge of our program.

A year ago, the Commission undertook a study on the validity of testing in education and employment. It focuses primarily on mental tests, including intelligence tests, achievement tests, and aptitude tests.

Many different areas of testing are covered—testing used in elementary and secondary schools, for admissions to higher education, for scholarship awards, for screening, hiring or promoting employees or for occupational licensing. The study is divided into three parts. Today we will discuss general issues concerning test construction. At some future time, we will address the appropriate uses of tests, first in education and then in employment.

Today we will talk about how bias is defined, what test makers can do to make sure tests or test questions are not biased, what procedures are required to validate tests, whether test developers should be monitored, and what should be done about the adverse impact of tests. We will also talk about related legal issues.

Let me introduce our experts.

Dr. James Loewen holds a degree in sociology from Harvard University and taught for several years at Tougaloo College. He is now professor of sociology at the University of Vermont. He has written extensively about civil rights issues in testing and education.

Dr. Nancy Cole is representing the Educational Testing Service, known as ETS. ETS is a major developer of educational and occupational tests, notably the Scholastic Aptitude Test. Dr. Cole has long been recognized for her contributions to the field of testing and has recently joined ETS as executive vice president.

Dr. Lloyd Bond is from the University of North Carolina. Until recently, Dr. Bond has been affiliated with the Learning Research and Development Center at the University of Pittsburgh. He has been studying the thought processes of black and disadvantaged test takers to understand why they have difficulty giving the correct answers to test questions. He has served on the board of trustees of the College Board, and in many other capacities concerned with testing.

FairTest, the National Center for Fair and Open Testing, is an advocacy group concerned with the issues we are addressing today. They were unable to be with us but Dr. Rudert, from our staff, will read their statement.

Ms. Wigdor has joined us today. She is with the National Research Council of the National Academy of Sciences. She was coauthor of their study of ability testing completed 7 years ago, and of a study of fairness in employment testing released 3 weeks ago. She will briefly describe this study to us.

Mr. Clint Bolick is the director of the Landmark Center for Civil Rights. Clint has worked with civil rights at the U.S. Department of Justice and the Equal Employment Opportunity Commission. His foundation is representing the parents of a black student who was denied the opportunity to take an IQ test in California.

Mr. Barry Goldstein is with the NAACP Legal Defense and Educational Fund, Inc. Mr. Goldstein has litigated many employment, discrimination, and other civil rights cases and is a frequent lecturer on these issues.

Today's record is part of the Commission's inquiry into the validity of test construction and use. Each panelist will have 15 minutes to make his or her statement. Drs. Loewen, Cole, Bond, and Wigdor will begin. Dr. Rudert will read the FairTest statement. This presentation will be followed by an hour of dialogue concerning test construction issues. Members of the second panel will be here by that time.

Let us begin then with Dr. Loewen.

Presentation of James W. Loewen, Ph.D

DR. LOEWEN: Good morning. Because the topic—the validity of testing in education and employment—is so broad, I will confine my remarks to testing in education and to the Scholastic Aptitude Test (SAT) in particular. My remarks, however, will be relevant to other education and employment tests that ETS develops.

Testing is an important civil rights issue because racial, ethnic, and gender groups differ in their test scores. African Americans score 170 points lower on the SAT (combined math and verbal scores) than whites. Hispanics and Native Americans also score much lower than whites. Women score 57 points lower than men. Rural students at the University of Vermont score about 200 points lower than students from suburban areas. Finally, at least among whites, SAT scores are lower for students with low parental incomes.

ETS claims that the test only shows that society provides far better education for affluent, suburban whites than for inner-city blacks or rural Native Americans. But test results channel students' college choices, determine their chance for financial aid, and affect their perception of their own aptitude.

On the Preliminary Scholastic Aptitude Test, the smallest difference—the 57 point gap separating women from men—causes two-thirds of all National Merit Scholarships to go to boys. This gender gap also determines who gets State merit scholarships, who participates in programs for gifted high school students, and who is admitted to some prestigious colleges.

Women lose out.

SAT scores correlate strongly with socioeconomic advantage. "Aptitude" testing completes a vicious cycle-socioeconomic advantage begets aptitude, otherwise known as a high SAT score, which then begets socioeconomic advantage.

Teaching experience has shown me that based upon their SAT and GRE scores, blacks think they have no chance at graduate school and women expect to fail statistics. Conversely, advantaged whites showing only a modest ability to read, write, and think, test well, and win admission to prestigious schools and fine jobs. Unless counter-balanced by robust affirmative action programs, standardized tests block the dreams of many minorities and women for equal access to education and employment.

Furthermore, test scores are biased. One-third of the gender gap on the math exam,[1] all of the gender gap on the verbal exam, and perhaps 40 percent of the black/white gap on the verbal exam, is due to test bias.

Issues of test bias and equal opportunity are intertwined with methodological issues like how best to develop test items, assess their validity, and examine them for adverse impact.

What do we mean by valid? "Valid" means that tests test what they claim to test (i.e., content validity) and correlate strongly with performance in college (predictive validity).

The process of writing valid test items is formidable but manageable. Four steps will handle adverse impact:

Step one: Write an item. Make sure that it tests skills or knowledge that a high school student should know, and decide how this item fits into the array of skills or knowledge that relate to college performance. That is face validity.

Step two: Share the item with culturally diverse referees. These referees should judge items by asking four questions:

a. Does the item contain offensive language? Does it present stereotypes?
b. Does the item use language that has different meanings for different groups? For instance, whites usually use the word "environment" to mean the natural environment, while blacks usually refer to the social environment. Both usages are correct, but an analogy based on the former will trip up blacks, while an analogy based on the latter will confuse whites.
c. Is the item unfairly unfamiliar to certain groups? Take, for example, this item from a recent SAT: "Oarsman is to regatta as . . . ." The reasoning in this item was elementary, but the item's vocabulary was more available to affluent eastern whites than to rural students, blacks and possibly other groups.
d. If the item does unfairly draw on one subculture, then does the test also include items that draw on other vocabularies to achieve balance?

Step 3: After the referees have approved the item, try the item in a test. Check for adverse impact by comparing the results—the percentage correct—by race, sex, income group, rural versus urban, and region. Drop items that markedly favor one group.

---

[1] Thus, for example, Loewen suggests that one-third of the average difference between blacks and whites on the math portion of the SAT could be eliminated if sources of bias were removed from the test. (Ed.)

Step four: Correlate the item with an output measure. For example, examine students' first year college grades to see if those who answered the item correctly did better than those who missed it. Again, this research should be done within each race, sex, and so on, and for the sample as a whole.

The first two steps examine the item's content validity and review its content for possible bias. The third and fourth steps use test results to check empirically for adverse impact in the item and to obtain its predictive validity.

What steps does ETS take to examine an item and build a test? ETS does step one. Test writers write items.

In step two, they share the item with a review panel that usually includes at least one woman and one minority person. ETS referees strike offensive language or stereotypes. This is the first part of step two. But it appears that ETS ignores the other three parts of step two.

ETS researchers have proven that some groups use a word in one way while other groups use it in another. Yet, ETS has never dropped or included a single item as a result of this research.

ETS does not use review panels to see if items might be unfairly unfamiliar to certain groups. Phyllis Rosser, John Katzman, and I examined the performance of 1,112 students on the SAT. We found 17 items that favored one sex or the other by more than 10 percent. Some of these items obviously favored males. No panel reviewing items for gender bias would have passed them.

Does ETS try to balance the test so that it is culture fair? ETS has considered the issue of overall balance on the verbal exam. In the 1950s and 1960s girls did better on the SAT verbal; boys did better on the math. In 1967, for instance, women averaged five points higher than men on the verbal. Around 1972, however, females lost their verbal lead as a result of ETS' changes in the content of test items. ETS changed the test to create "a better balance for the scores between the sexes." So today men's verbal scores average about 10 points higher than women. Is that a better balance? I think it is an outrage. I know of no justification for it.

Using existing SAT items, test makers could make an SAT verbal test on which women scored 50 points higher than men. They could also make a verbal test on which men scored 50 points higher than women. The 10-point gender gap which now exists is arbitrary and should be cut to zero. On IQ tests that was done long ago.

Even if ETS did review items adequately for content bias, content refereeing alone is inadequate to determine item bias. Referees cannot always detect biased content because some items favor one group for no obvious reasons. Items must be tested empirically.

This brings us to step three. The surest way to locate items that have differential impact is to compare the actual performance of different groups on the items. If an item markedly favors one race, sex, or group, then it should be removed even if the test maker does not fully understand the source of its bias.

ETS does not test items for differential impact, or at least it didn't through 1987, nor does ETS then remove items that favor one race, one sex, or another group.

Contrary to their claim, even if items with differential impact are removed, plenty of items will remain, and on all levels of difficulty. My paper will establish that.

Step four, predictive validity, or correlating the item with an output measure, is particularly persuasive of a test's validity. It is the best single measure of validity. ETS does not use it.

How does ETS check its items? ETS uses two statistics. The first is "point biserial correlations." Point biserial correlations actually increase test bias and adverse impact. For example, imagine an item on which blacks do better than whites. In testing jargon, such an item misbehaves. Its point biserial correlation will be lower, even negative. So it will never graduate from the experimental section to the real SAT. Neither will math items that favor girls nor any items favoring rural students or Hispanics. Thus, the point biserial correlation coefficient maintains a bias in favor of the status quo on tests.

Even among whites, the point biserial correlation is biased against those who live near blacks, Hispanics, or Native Americans. Within white America, white students with the most familiarity with black culture are those who attend inner-city schools or truly desegregated schools. Since the SAT is not multicultural, it ironically rewards white students in overwhelmingly white suburbs for knowing only the white subculture.

The second procedure that ETS uses to screen items is "differential item functioning." Differential item functioning (DIF) methods are intrinsically flawed. They remove the mean percentage difference before looking at the items. These percentage differences are the best measure of adverse impact. For example, on the math SAT we analyzed, only one math item had any verbal content that related to girls. That content consisted solely of the proper noun "Judy" in a problem: "Judy doubles K and adds 12." On that item, girls did well, only a half-percent below boys. By contrast, on an item set in a boys' camp, boys out performed girls by 12.3 percent. ETS' DIF procedures are more likely to flag the "Judy" item as biased towards women, than to flag the boys' camp item. The "Judy" item, on which boys and girls performed nearly the same, might be removed as biased while the boys' camp item would stay on the test.

Constucting valid tests makes tests more fair. If ETS and the rest of the industry constructed tests along the lines presented above, it would not provide equal opportunity but would give a closer approximation.

ETS will complain that to construct valid tests along my lines costs too much. Are there inexpensive alternatives? Yes, a cheaper alternative would simply remove the items with the most disparate impact. That is the Golden Rule procedure using percentage differences. Percentage differences should be used, not blindly, but as the best starting point. They correspond most closely to the meaning of adverse impact.

Another inexpensive approach is mean-balancing, which is similar to "within-group scoring" but conveys a different symbolic meaning. This method would add to the scores of low scoring groups the difference between that group's average and the white male average.

The industry can not be trusted to police itself. ETS won't change its procedures without Federal oversight. It's not in their interest. Test makers, even those with nonprofit status, are in business to make money. Research is not high on their agenda, particularly when it is expensive and might question their past testing procedures or leave them open to charges of bias or incompetence. I think this is the source of ETS's stubborn refusal to respond to criticism in the past.

This country needs competent, unbiased testing to provide equal opportunity and an efficient talent search for excellence. Federal oversight is overdue.

Presentation of Nancy S. Cole, Ph.D

DR. COLE: ETS invests enormous resources in ensuring that its tests are valid and fair. It has led the testing industry in implementing procedures to ensure fairness. Its researchers have produced much of the test data used by both its critics and its defenders. ETS welcomes inquiries into the integrity of its test development procedures.

What does it mean to say that a test is valid or that it's biased? Validity is having sufficient information about a test to show that it is appropriate for a particular use. Fairness is a part of validity. To be valid, a test must be fair. Bias is the opposite of fairness. It's a type of invalidity—invalidity or differential validity with respect to particular groups.

Validity must refer to a specific use. A test cannot be valid or fair in general, for all uses. Validity and fairness or invalidity and bias depend on the use. Dr. Loewen's discussion is clouded because he speaks generally rather than in relation to a particular use. For example, a spelling test might be valid for hiring secretaries whose work involves spelling, but the same test would not be valid for hiring janitors whose work does not require spelling. We have to ask "Valid for what?" before we can address validity or fairness.

Validity and fairness cannot be represented by a single number from a single approach. They are judged using five types of information: the use to be made of the test, test content and format, administration and scoring, internal test structure, and external test relations. Each is considered in developing a test.

First, the context of the test's use indicates what questions about validity are necessary. Tests can be used in various ways. Different uses raise different questions about validity. To ask the right validity questions, one must understand the context, with whom the test is to be used, under what conditions and for what purpose, what action is to be taken on the basis of the score, etc.

The second area concerns the appropriateness of the test content and format of the questions for the interpretations of scores. For example, one explores the areas of math included on a math test and the form of the questions used. This category includes "content validity," that is, obtaining expert judgment about the content domain. It also includes evidence about the appropriateness of the content for various groups—the fairness issue. For example, at ETS, test content receives a sensitivity review in which trained reviewers look for offensive content, stereotyping of groups, balanced references to different groups, when appropriate, and other content related to fairness issues. These considerations provide important evidence about validity and fairness.

The third area involves the way a test is given and scored—important factors in what a score means and whether it shows bias. There are concerns of "standardization," meaning giving and scoring the tests so all test takers are treated in the same way. Procedures must be consistent with the intended meaning of the scores. Many types of information are sought to check that the procedures produce the intended meaning and are comparable and fair for all examinees.

The fourth area is internal test structure. If a test and questions on a test are intended to have a particular meaning, then that meaning implies certain relationships among test parts. For example, a math test might include a total score as well as subscores in problem solving and computation. A question that's part of the problem solving section should be more highly related to problem solving than to the computation score. Each question should be highly related to the total score, and to a measure of the characteristic being measured by that total score. (These are where the biserial correlations come in.) These issues are addressed in internal test structure analyses. The purpose is to see if intended and expected relationships and properties exist. The widely discussed methods to examine possible item bias fall into this category of information, too.

An important part of the information about a test's validity and fairness for a particular use concerns how test scores are related to measures external to the test. This is the fifth category. For example, if a test is supposed to measure preparation for college work, then the scores should relate to eventual college performance. This relationship is called "predictive validity." In the context of fairness or bias, these questions involve the relationship of test scores with external variables like college performance for special groups, such as women or minorities.

If Dr. Loewen was suggesting that we use these predictive relationships during test development, we couldn't, of course, examine them with the test takers until a couple years after the test is given. So, his proposed procedures cannot be readily applied during test development. Instead, we regularly examine different types of items for these relationships to learn, for future test development, the kinds of items that are appropriate or inappropriate.

To recap, first, to address validity, we have to know validity for what. A test can be valid for one use and not for another. Second, we must consider many types of information. An answer to validity or fairness issues is not found in only one type of information or one single number. One must review a wide range of information to judge whether the evidence of validity and fairness is sufficient to support a particular test interpretation.

In spite of my exhortation about considering a wide range of information, discussions of bias have focused mostly on two types: indications of differential predictions for different groups and of differential performance on individual test questions. For example, the SAT's prediction of college grades in minority and gender groups has received a lot of attention. The strength of the relationship is at issue, as well as over or underprediction for some groups. I'll leave this area to the later paper.

Dr. Loewen and recent newspaper articles have referred to group differences in performance on test questions. Some complications arise in trying to make interpretations of group differences. First, raw differences between groups on test questions are meaningless for judging fairness. They're important for other reasons, but they do not clarify issues of bias. Second, there are better procedures for examining test questions that control for valid group differences. However, even with these controls, the judgments about whether or not to eliminate an individual question from a test remain difficult ones.

Differences between subgroups on important academic accomplishments concern all of us. We would hope that our educational and social system could produce opportunities that lead to equal performances by minority and majority groups, by males and females. To assume that differences could exist, as they surely could, may cause discomfort. Such differences may indict our social and educational systems. Some fear that the indictment will turn toward the lower scoring groups rather than the system. However, concern about such a difference is very different from concluding that the difference indicates that the question or test is biased.

Interpreting raw group differences on a question as bias rules out that the groups might validly differ. Suppose the test in question is a ruler to measure height of males and females. When the results showed that males tended to be taller, would we conclude that the ruler was biased? Or, to be considered unbiased, would we require that a test of Spanish fluency produce identical scores for native speakers of English and Spanish? To require equivalences between groups without regard to possible valid score differences is foolish, as these extreme examples show. It's not reasonable to assume that all groups will score precisely the same on every test, even though our social concerns might lead us to wish this were the case.

Consider the two SAT questions that produced the largest differences between males and females on the math and verbal subsections when the test was administered in New York in November 1988. On the verbal item, the difference between the percents of males and females giving the correct answer was 15 percent. The math item had a difference of 17 percent. Both items favored males. Should we assume that male and female test takers in New York are equivalent in verbal and mathematical reasoning skills and, therefore, conclude that these items are biased? Many people make this assumption and interpret such results as bias.

However, many educationally relevant characteristics are different for males and females. In our data from the SAT, 34 percent of the males and only 11 percent of the females planned to study physical sciences in college. Sixty-three percent of the males taking the SAT report having 4 or more years of mathematics in high school, whereas only 53 percent of the females do. Would we expect groups with such differences to score the same on mathematics? In addition, more females than males take the SAT. This suggests that the particular self-selected males and females taking the SAT are not equivalent representatives of their gender groups.

To conclude that the scores of groups differing in many ways should be the same is to expect the unexpected. We must recognize the possibility that groups may differ validly on test scores and not interpret such differences as bias. If we accept that males and females, or other groups, might validly differ, then we're forced to reject raw group differences as evidence of bias. Instead, we look for ways to control for the possibility of valid differences.

The general approach has been to seek test questions that yield larger or smaller differences between groups than the test as a whole. This is a common way technical scholars have proposed to look for anomalous questions. Note that this is a search for anomalous differences, not necessarily bias. At least some of the anomalies that appear from such analyses might be due to bias. ETS has implemented one of this class of analyses. The procedure is intended to identify differential item functioning for groups. We call the procedure "DIF." DIF provides statistics that show when an item is operating differently for different groups after controlling for overall differences in test scores.

We use the DIF results to sort items into three classes. Items labeled "A" are those that do not show anomalous behavior. Items labeled "B" are those that show minimal, modest differences. Items labeled "C" are those that show rather substantial anomalous behavior. We perform DIF analyses to compare gender groups and compare black, Hispanic, Asian, and Native American groups with white students.

When pretests have enough students in different groups, we do these analyses on pretest data. Our rules for test assembly call for use of "A" items before the "B" items and the use of "C" items only if necessary to meet content requirements. Typically "C" questions are eliminated from further use at the pretest stage. For tests for which we're unable to perform such analyses at pretest, the analyses are run after an actual test administration but before the scores are reported. "C" items receive the most attention and may be eliminated before the scoring is done.

On the two sample SAT questions, DIF analysis produced very different results. It rated the math question as an "A" question. This analysis shows that when we take valid differences between males and females on math into account, the differences on this question are not statistically or practically significant. The question is not anomalous in relation to the rest of the test. It functions like most of the rest of the questions on the test. Any claim that the question is biased is arguable.

The verbal question received a "C" rating, raising the more interesting set of issues. Our first step with such a question is to see if we can understand the source of the anomalous statistical behavior, that is, the characteristics of the question that produced the "C."

For example, occasionally the content is stereotypically associated with one group more than another in ways unrelated to what is being measured and might therefore produce the anomaly. In comparing males and females, questions using sports examples sometimes receive "C" ratings even though knowledge of the sport is not required to answer the question. In such cases, we typically seek a replacement question. Among verbal questions, we occasionally find a word that is more familiar to one group than another. The "regatta" item—to which critics frequently refer, although it is not a recent item—is exactly the kind of item we'd expect to show up on the DIF analyses. Recently, a question using the word "mink" referring to an animal, not a fur coat, was flagged by the DIF procedure during pretest as a "C" question in contrasting performance of black and white students. That question was replaced.

Other questions, such as the verbal one administered in New York, pose a different situation because the test developers and sensitivity reviewers who reviewed the question could not identify an aspect of the content to account for the anomaly. We're left to wonder if there's a content problem that we're not able to identify, if the "C" rating could be unique to this group of test takers and not an enduring characteristic of the question, or if this anomaly is not related to our concerns about unfairness.

It's not reasonable to conclude that a question or test is biased on the basis of raw group differences. Such differences are of great concern to us as citizens and educators, but to require that every test and every test question produce identical results between every group is absurd. Groups differ on many characteristics and they will probably also differ in many skills that tests measure. Thus, to account for this, we need to look for differences in questions other than differences produced by overall valid group differences.

An appropriate way to search for anomalous items is to control for group differences on the total test score. The DIF procedure used by ETS is one such procedure. However, the examples illustrate that the judgments about bias are very difficult even after the DIF analysis.

Our society's educational and social problems are difficult. All children do not receive comparable educations in our schools. Their homes do not provide them equal starts on that education. Even the same homes do not necessarily provide boys and girls equal starts on that education. Our social institutions do not serve us all equally well. To hide the effects of such inequalities as test bias is a foolish, and potentially dangerous, self-deception.

At the same time, we must have high standards for test quality when tests are used for important decisions about human beings. Extensive and complex analyses of test fairness and validity are required. These analyses will not often yield a simple answer. Furthermore, tests may have substantial validity for one use and little for another. We must use the best professional expertise and judgment about many types of evidence to conclude whether a particular test has adequate validity and fairness for a particular interpretation to be used in a particular situation.

Presentation of Lloyd Bond, Ph.D

DR. BOND: The first two presentations highlighted a controversy. Dr. Loewen seems to think adverse impact and bias are identical concepts. Dr. Cole thinks they are fundamentally different concepts. I am convinced that they are different concepts. The simple observation of score differences between males and females, or blacks and whites, rural and urban children, is by itself not sufficient for showing bias.

Once I was summarizing my work with black high school students who were doing extremely well in high school math but poorly on the SAT. An ETS researcher asked if we could somehow change the content, that is to say, the context of math items to remove the differences between black and white youngsters. I thought of an item like the following: "This black family was travelling at a speed of 15 miles an hour . . . ." How is that going to help? Surface attempts to change items in order to overcome miseducation is very misguided.

About a year and a half ago I began watching students try to solve problems on the mathematics section of the SAT. This research has convinced me that, at least in math, the differences between boys and girls, and blacks and whites, represent real differences in achievement. We have to address that issue rather than trying to ascribe it to some inherent fault in standardized tests.

The verbal section is an entirely different matter.

I had short responses to your questions. How should biased items be identified? I don't know.

Should biased items be categorically eliminated? If they are biased, yes, eliminate them. But, bias and adverse impact are fundamentally different.

What proportion of items in current tests are biased? I don't know. How much does eliminating items with DIF reduce group differences? Differential item functioning per se has not reduced group differences that much, but there is a distinction between DIF and bias and adverse impact. I hope to elaborate on these differences in my paper.

Is there differential predictive validity for black, white, male and female? I think not but I'm not sure on that point either.

How high should correlations be for a test to be valid? A test may have a very low validity coefficient and still be useful for some purposes.

If predictive validity is high across groups, is it necessary to obtain other forms of validity?

Yes.

How should job analysis and content validation be done? Very carefully.

Written Statement Provided by FairTest

The National Center for Fair and Open Testing (FairTest) is dedicated to ensuring that the more than 200 million standardized multiple choice tests administered in the U.S. each year are fair, open, valid and relevant. Unfortunately, many current exams fall short of these minimum standards. Since 1985, FairTest has been publicizing the shortcomings of these instruments and urging reforms in tests and their uses.

This testimony will discuss, first, the inadequate test validity that plagues test use in elementary and secondary education, college and university admissions and employment; and, second, FairTest's testing guidelines.

Validity in standardized tests tells us whether a test measures what it claims to measure, how well it measures it and what can be inferred from that measurement. Test validity cannot be measured in the abstract but only in the context of specific uses of test results. Thus, information and conclusions regarding test validity in one context may not be relevant and applicable in a different context. A test that does not measure what it claims to measure is not only invalid, it can be dangerous.

Construct validity should be the underpinning of validity in educational testing. For a test to have construct validity, it must adequately measure the underlying theoretical trait it claims to measure. For example, does the test accurately measure "academic potential" or "competence" or "reading"? To answer such questions requires an accurate grasp (construct) of the trait to be measured (for instance, "reading") and knowledge of how the test scores will be used.

Many tests lack construct validity, that is, they do not measure what they claim to measure. For example, a test that is used to make statements about school achievement may really measure another construct such as "verbal ability." In part, this is because the multiple choice format is limited. For example, writing is not selecting a missing word from among four or five choices to insert into a sentence or finding errors in a text. Yet many tests measure writing ability in this way. While the multiple choice format can measure knowledge of simple information, it generally cannot assess the ability to use or create knowledge, though test results are often used as if that ability is measured.

The dangers of inadequate construct validity are two-fold. First, if the test measures something other than what users think it measures and is used in selection, low scorers who can perform well may be excluded or high scorers who cannot perform well may be included. Second, use of tests with inadequate construct validity may result in improper and misdirected teaching. For example, this occurs in a reading class where students fill out mimeographed worksheets that simulate multiple choice standardized tests instead of reading, discussing and writing. Pressures for good performances on standardized tests can inappropriately drive curriculum and teaching.

Problems also exist with predictive validity. In education, predictive validity is sometimes established by the rather circular procedure of comparing results on one test with results on a second without establishing just what the second test measures. At other times, tests are validated by comparing scores with teachers' grades. This begs the question of how to determine which is more valid when the results diverge.

Tests for young children reveal further problems with predictive validity. Not only do I.Q. and readiness tests lack adequate constructs for "intelligence" or "school readiness," they often measure little more than social background. Despite these flaws, test scores are used to place and track young students in "dumbed-down" classes which lead to inferior education and create a self-fulfilling prophecy.

The recent controversy surrounding the NCAA's Proposition 42 further illustrates the limitations of predictive validity and the dangers of misinterpreting validity. Proposition 42 will, if implemented, bar many colleges from giving athletic scholarships to students who obtain less than a 700 on the combined SAT or 15 on the ACT. Such test use assumes that students who score below an arbitrary cutoff point cannot do college level work. This is a predictive claim. However, a study by Dr. Timothy Walter at the University of Michigan found that 86 percent of those who would have been barred under the rule did acceptable freshman level work, which is all the SAT and the ACT claim to predict. In fact, no predictive validity study exists to support the view that those who score under 700 or 15 cannot do college level work.

A similar situation exists with the National Teacher Exam (NTE) and other tests that prospective teachers in many states must pass to be certified. The NTE claims to be a minimum competency test. It does not claim that those who pass will be good teachers, only that those who fail cannot be good teachers. But no study has proven that those who fail would be disproportionately poor teachers. In fact, counter examples exist. Last year in Prince Georges County, Maryland, provisional teachers whose supervisors rated them satisfactory or better were not hired as permanent teachers because they did not pass the NTE.

Low correlations between tests and job performance are common. The typical correlations of from 0.2 to 0.4 mean that test scores explain from 4 to 16 percent of the observed difference in performance. Such results provide insufficient explanation or prediction of worker success to warrant making decisions solely, or even primarily, by test scores.

As with most other tests normed on the majority population, minorities score lower on most employment tests. However, the low test scores of minorities often do not predict job achievement. For example, results on many administrations of the GATB, the General Aptitude Test Battery, have been compared with supervisor ratings. High performing blacks frequently score lower on the GATB than low performing whites. This occurs despite known problems of racial bias in supervisor ratings.

Similarly, the SAT shows differential prediction for men, whose college performance is overpredicted, and women, whose performance is underpredicted. Because the degree to which women are underpredicted is less than that to which men are overpredicted, the makers of the SAT claim that the SAT is a more "valid" predictor for women than men. But this claim diverts attention from the real issue: college entrance and scholarships often hinge on total test scores. The SAT gives men an unfair advantage.

With respect to bias, first, a biased test is an invalid test. Second, the problem of bias is not just one of detecting biased items, but of appropriately assessing people from a variety of cultures. Third, group differentials on test scores ought not harm a lower scoring group unless the scores can be proven to accurately predict future performance. Finally, in the case of education, use of test scores must not result in some groups receiving an inferior education.

FairTest believes that when properly constructed, validated and used, standardized tests can serve as a useful though limited tool in assessment. However, it has become all too obvious that standardized tests often are not properly constructed or validated. Moreover, their misuse is creating problems for students, teachers, schools and university and employment applications. The question arises, then, what should be done to reform tests and test use?

Reflecting its concern over the misuse of standardized tests in U.S. society, FairTest's Test Reform Agenda is guided by four principles:

First, tests must be properly constructed, validated and administered. Tests should measure pertinent, not extraneous knowledge differences among students or applicants. Questions must be relevant to the knowledge, abilities or skills being tested. Test items and instruction should be written clearly and accurately.

The tests themselves should take into account the diversity of language, experience and perspective embodied in the test-taking population. At the same time, questions and scoring procedures should acknowledge the complexity and diversity of intelligence and individual development.

Test validation should ensure that the content of the test matches the content of what is taught or done on the job. But test developers cannot stop at content validation. They must document assumptions about the relationship between test results and future performance. At the same time, they must demonstrate that test results are accurately related to the underlying knowledge, skills and abilities the test claims to measure.

Second, tests should be open. Public schools, test takers and independent researchers should have access to the descriptive and statistical data needed to verify test publishers' claims regarding test construction and validation. This should include the release of questions used on previous tests as well as data on test results identified by race, ethnicity, gender, socioeconomic status, geographical residence and other demographic distinctions.

Publishers also should release information on test construction and validation. Test users or independent public agencies should be able to investigate the claims of test publishers regarding the construction and validity of the tests. At the same time, users should disclose and monitor their own process for test administration and guidelines for test use.

Third, tests should be viewed in the proper perspective. Both test developers and test users should work to ensure that test results are properly interpreted and employed by schools, colleges and universities, employers, policymakers, test takers and the general public. As the 1974 *Standards for Educational and Psychological Tests* state, "A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances." Test users too often ignore this statement. At a minimum, test scores should not be the sole or primary factor in educational or employment decisions.

Test developers and test users must recognize that standardized tests are only limited measures of educational reality. Used alone, they distort what they seek to measure, and often undermine the quality of education offered in our public schools. Both test developers and test users have the affirmative obligation to promote a proper, reasonable and limited use of standardized tests as one of a series of assessment mechanisms.

Fourth, appropriate and authentic assessment instruments should be used instead of standardized tests whenever possible. Standardized multiple choice tests can only measure a very limited range of knowledge, abilities and skills. New technologies and a better understanding of learning provide opportunities to measure them more fully and accurately. Educators and employers should invest in developing and using new methods. They can be used to diagnose the strengths and weaknesses of students, to help them learn, rather than to sort, stratify or segregate them. And more accurate assessment of college and job applicants can help both applicant and institution.

Although FairTest believes that institutions that develop and use standardized tests have the primary obligation to reform tests and test use, the government has a role, too. By establishing guidelines for the testing industry, requiring information on standardized tests to be made public, and analyzing test results to guard against bias, the government can improve the quality of tests and test use. More importantly, public agencies can set the standard for proper use of test results. Too often, government is the biggest misuser of standardized test results.

Unfortunately, too many policymakers and educators have ignored the complexities of testing issues and the obvious limitations they place upon standardized test use. Instead, they have been seduced by the promise of simplicity and objectivity. For this infatuation with tests, our people have paid a high price in damage to schools and employment opportunities and in the loss of social equity. Unless Americans act now to limit and reform the use of standardized tests, that price will continue to increase.

Comments of Alexandra Wigdor

MS. WIGDOR: Rather than talk about test construction, I am here to describe a recently published report from the National Academy of Sciences, National Research Council. This report is called, *Fairness in Employment Testing.* It is about the widely used employment test, the General Aptitude Test Battery (GATB).

The GATB is a general test of cognitive, perceptual, and psychomotor skills used to predict job performance. It was developed by the Department of Labor in the 1940s and, in the last 40 years, has been used in the Public Employment Service. Every county or town has a job service office which helps match job seekers and employers. This test might be used for placement in some jobs handled by that job service office.

In 1980 the Department of Labor began a new experimental use of the GATB. New developments in measurement practice and statistical theory in the last 20 years encouraged the Department of Labor to promote use of the GATB to refer people, not just to the 500 jobs for which validity studies have been conducted, but to all jobs.

The theoretical field which allowed this new use of the test is called meta-analysis, or in this testing field, validity generalization (VG). The theory of validity generalization provides formal rules for extending the results of test research. With these procedures one can estimate the validities of a test for performance on new jobs based upon the validities for jobs already studied.

The General Aptitude Test Battery has been validated for some 500 jobs over the last 40 years. However, the U.S. economy has more than 12,000 jobs. The question is, "can this test which has a certain degree of validity for 500, jobs be assumed to be valid for 1,000, 5,000, or all 12,000 jobs?"

In 1980 Department of Labor research contracts provided optimistic estimates of validities of the GATB for all 12,000 jobs in the U.S. economy. Consequently, the Department of Labor encouraged the Employment Service to start using the test much more widely. And to give employers the maximum economic benefit of testing, the Department promoted a system of referral based on ranking by test score (rather than, say, a minimum competency referral system).

Because the Department of Labor is concerned with the problem of adverse impact, it introduced a within-group percentile scoring system when promoting this new system. Within-group scoring computes the scores of blacks, Hispanics, and all others according to percentiles within their own group. This scoring procedure simply eliminates the difference in mean (average) scores among the groups. For example, within the black group, a score of, say, 235 might fall at the 50th percentile. The 50th percentile in the white group might be 280. When you convert to percentile scores within groups, blacks and whites, at their respective 50th percentile, are referred at the same time even though their scores were very different to begin with. The Department of Labor introduced this system of computing scores for blacks, Hispanics, and others to answer two important social needs: First, the Department wished to comply with its understanding of equal employment opportunity, and second, to provide employers jobseekers with the highest predicted job performance.

In 1986 the Justice Department found out about the scoring system. Mr. Reynolds, the then-Assistant Attorney General for Civil Rights, informed the head of the U.S. Employment Service in the Department of Labor that, in his opinion, the use of the test with these score adjustments was illegal and unconstitutional. The two agencies felt an intensive study was warranted. However, until it was completed, they decided to maintain the status quo. The Department of Labor and the Public Employment Service would continue using the test in this new way where it had already been introduced, but would not introduce it in new offices. The Department of Justice would not issue cease and desist orders until a group of experts conducted a study.

This is the requested study. It has just been completed. It involved 2 years of extensive research and a re-analysis of all 700 studies on the GATB.

The study asked three basic questions. One, how good is the GATB? Is its intrinsic quality good enough for widespread use throughout the Employment Service? Two, what about validity generalization? Can the GATB be used for a much larger range of jobs than those in the actual validity studies? Three, what about score adjustments? Can scores be computed fairly and yet represent the employers' interest in getting the most efficient work force possible?

First, is the GATB good enough for this more ambitious use that the Department of Labor has envisioned? Our answer is a very qualified yes. The test is good as employment tests go, but no employment test is very good. Nothing is perfect.

We were a little bit surprised. The test is pretty old. It was first developed in the forties. After analysis, we came to the conclusion that despite its age it has about the same range of validities as other broad-based employment tests. We compared it to the Armed Services Vocational Aptitude Battery (ASVAB), a much more recent test with a much more ambitious development program. In comparison, the GATB does not look bad in reliability and validity. It also does not look perfect. It provides consistent measurement and is valid enough to be useful.

How valid is valid enough to be of some use? Research done for the Labor Department in 1980 estimated the GATB's validity as about 0.5. Our calculations are more conservative and less optimistic than this research. Our calculations from the 500 studies estimate the range of validities for the GATB as about 0.2 to 0.4, averaging 0.3. We think 0.3 is right.

Contrary to the FairTest statement, that figure is not to be dismissed out of hand. If the GATB had perfect prediction, it would have a 1.0 correlation, or 100 percent accuracy in prediction. A 0.3 correlation means you have about 30 percent of what you'd have if the predictions were perfect. On a scale of zero to 10, this is about a 3. (In fact, you will not find any test that is even a seven.) It's useful, but it's not perfect.

The next question is about validity generalization. The Committee found that, contrary to the general thrust of the Uniform Guidelines[2], this range of validities (0.2 to 0.4) would hold for a great many jobs in the U.S. economy. This finding does not mean that you can stop doing research. But, for the kinds of jobs that the Employment Service uses the GATB, one can reasonably assume that these validities will hold. This finding may cause policymakers in the Federal Government to rethink the meaning of the Guidelines.

Third is the question of within-group scoring. How do you compute scores within the context of civil rights laws and the concept of fairness? The study adds a scientific analysis to the more general fairness arguments.

If the test is useful but not perfectly valid, predictions contain errors. Thus, some people who get low scores on the test are not referred to employers and could have done well in the job. Conversely, some people who score well on the test, will do poorly on the job. That is the other kind of prediction error.

Figure 13–1 illustrates the point. Those who are predicted to do well, will indeed do well. Those who are predicted to do poorly, will indeed do poorly. Prediction error is found in sectors B and D. Particularly in sector D, those who do poorly on the test get low scores on the test and therefore would tend to be screened out and not referred to jobs, but nevertheless could do well on the job. We focused on the error in sector D in drawing our conclusions about computing scores.

This prediction error has nothing to do with test bias, but creates a problem when it's effect is combined with average group differences in scores.

The figure shows two ellipses representing the points where test score and performance score meet. The black group has a lower mean because blacks, on average, score lower. Therefore, proportionately more blacks fall into Sector D, the error field. Proportionately more blacks who fail the test will be able to do well on the job. Proportionately more whites do well on the test, but will not do well on the job. Other parts of the ellipses show the accuracy in prediction, apart from error. More blacks will do poorly in the test and would do poorly on the job. More whites would do well on the test and would do well on the job. So real group differences show up in this

---

[2] Uniform Guidelines on Employee Selection Procedures (1978), 29 C.F.R. Part 1607.

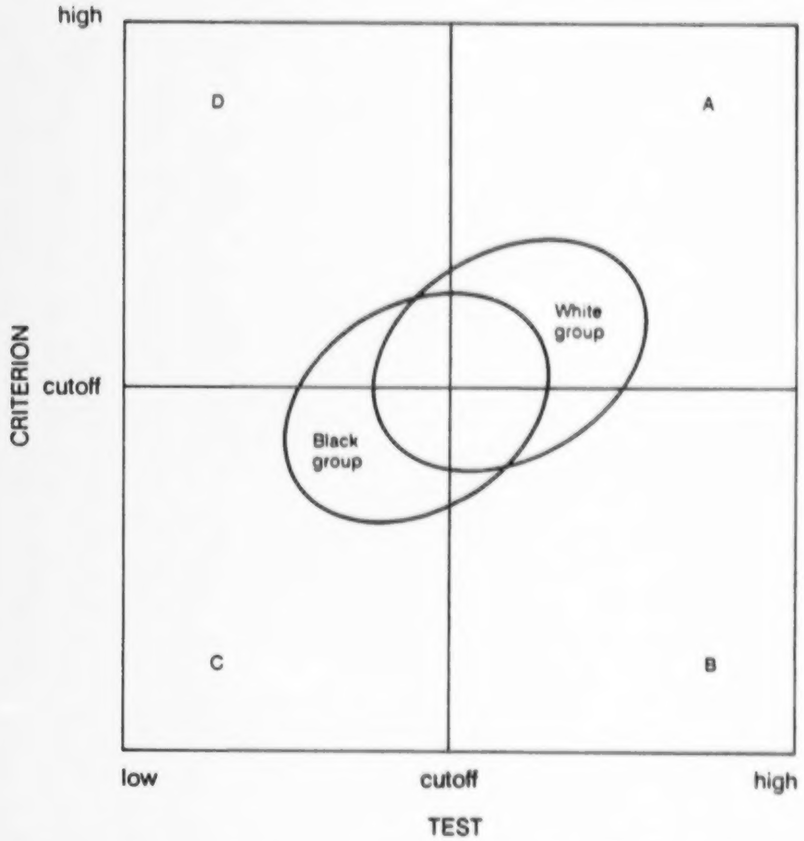
high
D
A
White group
CRITERION
cutoff
Black group
C
B
low
cutoff
high
TEST

FIGURE 13-1 Effects of imperfect prediction when there are subpopulation differences in average test scores.

validity research, along with error.

The Committee concluded that this error, which is not test bias but a combination of high and low scores and group differences, should not be allowed to disproportionately affect black and Hispanic jobseekers. We have therefore recommended that policymakers make score adjustments commensurate with the error of prediction in the tests. We have not recommended straight proportional referral of blacks, whites, Hispanics, or anybody else. We recommend making the adjustment commensurate with prediction error so that qualified people in all groups have the same probability of being referred.

Discussion

COMMISSIONER CHAN: Ms. Wigdor, how do Asians and Hispanics fare with the prediction error shown in your chart?

MS. WIGDOR: On this test, as on many other tests, there tends, on average, to be one standard deviation difference in scores between blacks and whites, and about half that much between Hispanics and whites. Thus, the Hispanic ellipse would fall somewhat above the black and somewhat below the white one—midway in between.

The Department of Labor has no separate data for Asian Americans, so we had no way of studying them. However, Asians surpass whites on many other tests.

COMMISSIONER GUESS: Ms. Wigdor, the overwhelming number of jobs in this country are still using training and experience to rank and employ candidates. The Civil Service uses these criteria extensively, as well. Does your study address the fairness of using training and experience to rank job candidates?

MS. WIGDOR: Since that wasn't part of our mandate, we didn't do scientific analyses of the validity of such criteria. Proven things like experience and education should be used in conjunction with, or to supplement, test scores. If the Department of Labor is going to use the GATB more widely, they must allow employers to apply other important selection criteria. Whenever you can supplement test scores with good information, you ought to do so.

No single criterion of selection is as good as multiple sources of information. We make that assumption in the report but we have not tried to measure the increment of using test scores in addition to, or instead of, other information. That wasn't part of our study.

COMMISSIONER BUCKLEY: Dr. Cole, your examples used items differing in concrete and abstract thinking. Mary Meeker shows that girls think differently than boys do. Girls think more abstractly and we are not teaching them to think in concrete terms. We need to teach the girls to use symbols better. On the other hand, we need to teach our boys to do abstract thinking.

If you identify test items as concrete or abstract, what racial, ethnic and gender differences occur? I'll bet differences occur in abstract and concrete thinking for ethnic groups and those differences would be very simple to overcome with training.

Are you looking at any of this research in assessing the bias of test items?

DR. COLE: I chose the most extreme examples, although I could present a long list of items.

ETS is very aware of differences between concrete and abstract thinking. The SAT is historically linked to more abstract thinking because of its role in college level work. For that reason, we focus the SAT on abstract thinking, reasoning, and problem solving in verbal and quantitative areas. However, abstract thinking may not be the right focus for every test and every purpose.

Some institutions use achievement tests along with the SAT because they focus more on particular content learning from school. That's an appropriate supplement. The SAT assesses only part of the preparation of children for college.

COMMISSIONER BUCKLEY: Dr. Bond, could you comment on differences in abstract and concrete thinking as evidenced in your work?

DR. BOND: I talked to 28 kids and I interviewed them extensively. Six were white; 15 were girls.

I did not find any *sex* differences in that aspect of the items, principally because I chose students who were doing well in school and poorly on the SAT. But the abstract/concrete distinction clearly affected performance.

One item went as follows: "At a certain college, one student consumes x liters of milk per month. At this rate, how many months will y liters of milk service z students?" No one could get the item right. I then changed that item to read, "At a certain college, one student consumes 3 liters of milk. At this rate, how long will 100 liters of milk service 4 students?" Everyone got it right. Even though the change in the item is superficial, its level of difficulty changed dramatically.

I am convinced that that is an instructional problem. There are probably also other matters—the amount of time that children spend on these things at home and whatnot. We as a nation can't control that. All we can control is what goes on during school.

COMMISSIONER DESTRO: Many of the tests seem to be tests of aptitudes, but I'm not sure that the people who use the test know what aptitudes they're measuring. Is the problem that the people who are actually accepting the students don't know what they want out of the test? Or, is it from the perspective of the testers? How important is construct validity?

DR. COLE: It's very important.

Construct validity is looking for information about what the test is measuring in the several areas I mentioned. In the past, we may have looked just at content or just at the prediction. Now, however, testing professionals realize we must understand what that test score means and doesn't mean. That's what construct validity addresses.

I agreed with FairTest's statement about construct validity.

DR. LOEWEN: I have a lot of problems with construct validity, and I disagree with FairTest's and ETS's emphasis on it. I don't understand it, and I think if we don't understand something, we should not accept it.

I think there are two kinds of validity—content validity and predictive validity. If a test, for instance, requires people to read something and then write something to show they learned what they read, that seems to be part of what you do in college. So, for college admissions, this test seems reasonable. It has content validity.

Predictive validity: Suppose a question reads, "What's your favorite color?" and somebody said, "Magenta." If everyone who answered, "Magenta," did "A" work in college and everyone who answered, "Purple," did "D" work, then that question, silly as it might seem, would have predictive validity. We'd have to respect its predictive validity, even though preferring magenta has nothing to do with the content of college work.

I don't think that the "A" in the SAT is merited. I don't think it should be called the Scholastic Aptitude Test. I don't think SAT scores measure who will be an apt student next year in college precisely enough to label a student as inept or apt.

DR. BOND: The controversy over the distinction between aptitude (or ability) and achievement is both ancient and continuing. Test developers themselves are unable to sort out the meaning of these two "constructs."

For example, one test is called the Otis-Lennon Mental Ability Test and another one is the Stanford Achievement Test. Researchers commingled items from these tests and asked testing experts throughout the world to distinguish the mental ability items from the achievement items. No one could.

I think I know what the distinction between achievement and aptitude is, but I couldn't put it in words.

DR. COLE: The word "aptitude" in the Scholastic Aptitude Test has existed for a very long time. The historical distinction between aptitude and achievement is whether the purpose of the use was to look forward and predict something or to look back and judge the accomplishment of something. "Aptitude" was associated with looking forward to predict something and "achievement" was to look back and judge accomplishments.

The word "aptitude" has become associated with intelligence or an inherent characteristic in an individual. This is not the intended meaning of the word, but because those associations are made and wrong interpretations follow, I would also prefer that the Scholastic Aptitude Test didn't have the term "aptitude" in its name.

DR. BOND: I agree.

COMMISSIONER RAMIREZ: My bias against tests came very early. I grew up in the lowest income school district in Texas. We were 97 percent Hispanic and 3 percent black, and the blacks spoke Spanish. I always scored high on tests, but knew that I was not smarter than the children that I was going to school with. I was 2 to 6 years younger than my classmates, but I was not smarter.

In high school I questioned the credibility of test results when I scored well on achievement and science tests, when I won the Betty Crocker Homemaker of Tomorrow Contest based on test scores, and when I scored the highest on the Armed Services test for mechanical ability.

I'm sure the tests have improved since then. However, my biases were confirmed when I taught students in that same school district. I am convinced that tests did not measure the potential in those students.

What's happening to kids at kindergarten, in first, second, and third grade? What's happening to them in ninth grade? What happens to them as they leave high school? Our public education system is in deep trouble. We are not succeeding in educating significant numbers of students. Presumably, the problem is mostly with what happens in the classroom. Teacher examinations, junior rising examinations, "preprofessional" tests, and SATs are used to predict and select the people who will best educate children. Yet, judging by the problems in the classrooms, the tests must be testing something other than what it takes to teach children effectively.

How can we constructively differentiate the valid tests from the invalid tests? When do testing professionals stop looking at research data from the perspective that formed it and go back to do basic research on what is really going on, whether it's in classrooms or in the minds of the subjects of your inquiry? Are we asking the right questions to begin with?

MS. WIGDOR: That's what construct validity does and why it is important.

DR. LOEWEN: Because test makers believe a new test item has to correlate with all the old test items, the test construction process builds in inertia, making change very difficult.

DR. COLE: A test is nothing but a sample of behavior of what a person can do right now on a particular set of questions.

The purpose of some settings is to change the status of that individual. The purpose of the educational system—schools—is to start with where a student is and see that s/he grows and improves in important skills. The student's status will change accordingly.

In other settings the purpose is not to change the status but to see what the status is and act accordingly. Such settings may have no mechanism for change. In personnel selection, the employer's purpose is to sort and select applicants for jobs. The employer may not be able to have individuals lacking skills catch up.

Higher education is complex. The purpose of some institutions and areas of higher education is to change the status, and in others the environment is very competitive.

The purpose of teacher licensing is not to change a teacher's status per se, although teacher status will gradually change over time with experience gained as a result of the license.

Sorting out different purposes in the context of the need for change, the intended change and the purpose of change will help resolve our problems with testing. Test scores should not create the expectation that the education system can do nothing for its children. That's exactly opposite to the effect tests should have.

DR. CUNNINGHAM: Once external validation is established, that is, you have a good fit between the test and a measure of performance, a test developer begins looking at internal methods of validation, looking at the test. The presentations of Dr. Cole and Loewen raise two very different ways of looking at internal validation. One controls for test differences (that is, overall ability levels or performance levels) in looking at item differences. The other does not. It looks at any differences in test questions. Is this a difference in the way we look at the available data? Can we reconcile those procedures?

DR. BOND: Currently, that is one of the most hotly contested controversies. Will we consider an item biased or flawed if groups differ in the proportion who can answer it correctly, or will we consider an item biased or flawed only if we have controlled for how the two groups did on the other items, and then compare their performance on the item? If, as many people believe, the test is categorically biased against certain groups, then any kind of internal analysis, like equating for performance on the other items, will not get at that kind of bias. Thus, there is a certain circularity to the argument.

I am offended by the notion that blacks are a different species and when we ask black youngsters questions in mathematics—to solve problems and to reason quantitatively—we are hopelessly and inevitably biased. (Verbal questions may be different.) This notion suggests that as black people we cannot respond to situations that others in the culture can, even though, as of about 1950 and after, every child in the country has encountered multiple-choice, standardized tests. These tests are part of our culture.

DR. COLE: The question of test bias is not a choice between these two procedures. The question of test bias involves looking at all the information, including the content of the test. Differences between groups give clues about extraneous content. But neither the DIF analyses nor overall group differences are automatic indicators of bias.

If there were bias throughout the whole test, yes, the DIF analyses will not show it. DIF analyses tell what items are anomalous or different from the way the rest of the test operates. These anomalies lead test makers to re-explore test content and what we learn from these analyses leads, in turn, to excluding some things from tests.

The more important analyses of bias have to do with prediction, with content, and with our own judgment, as educators, about the test questions. To throw out a test that shows group differences without looking at the test questions is silly. We must always look at the test questions, whether or not they show group differences. The questions must address important skills. The SAT must measure skills for college work; employment tests must measure skills relevant to the job. In making a test the most important focus is on test content—not group differences in test items—whether or not overall test performance is taken into account.

DR. LOEWEN: Earlier, two panelists suggested I do not understand that adverse impact and bias are not the same. I do. Bias is only one cause of adverse impact. I believe that bias causes probably all of the male-female difference on the verbal test and about one third of the male-female difference on the math test. Other things account for the rest of that male-female difference on the math test. A recent study indicates that differences in coursework and interests cause perhaps another third of the male-female difference on the math test. Perhaps the third third relates to societal expectations of girls versus boys in processes which are hard to name or identify.

I focus on bias because society is outrageously biased against women and minorities. This bias affects people by the time they're 17, then we test them with a test which has an additional bias built in. That would be the easiest bias to eradicate and reverse. Why shouldn't we include items on black culture so that whites who know things about black culture do better on them? And blacks of course will do better on them. But the opposite is going on.

Whether they are biased or not, tests have adverse impact. Test results indicate that there is unequal schooling in America, that we need better instruction in math and so on. But tests affect 17-year-olds. They affect where they go to school. They affect their self-perception and so on.

What do we do about the part of adverse impact that is not ETS's fault, that is not due to bias? Do we allow it to deny women National Merit Scholarships? Do we allow it to deny blacks and Hispanics admissions to college unless they come in by a stigmatizing affirmative action program? Or do we use some creative method such as adding mean differences to the score so that the sins of the past do not continue social inequality into the future?

VICE CHAIRMAN FRIEDMAN: I'm concerned about the politicization of the testing issues. The movement to change the testing system is more than just a need for objective information and to overcome bias. It is a movement to restructure who gets what in our society.

Politicization can be dangerous. Recently a Detroit orchestra abandoned the system of anonymous testing of musicians. Legislators had threatened to reduce funding unless some groups were better represented among the orchestra's members.

What role do politics play in this discussion of testing?

MS. WIGDOR: I would prefer to use the word "policy." These are policy issues: "Who gets what in society?" "How should opportunities be allocated?"

Social goals and necessities, like the competitiveness of this economy, must be balanced. Our recommended score adjustment is a policy recommendation in one sense. If the social goals are to optimize productivity and to bring as many minorities into the work force as we can, it's a policy recommendation. However, the recommendation is not politicized in the sense that it adjusts for high and low scores and the statistical effect that occurs when the technology is flawed. It doesn't have to do with the policy goals of blacks or whites or other groups.

We think the technology is probably useful, but equalize the negative effects of its flaws so that they fall the same on all groups. Focus not on test score but on performance. When you focus on performance and getting people at the same level of performance to have the same opportunity to be referred, that's equality.

Distinguish between policy and politicization and then distinguish at least for this report between the scientific arguments and the policy arguments. If we can disentangle those things, we can speak less heatedly.

DR. BOND: Disentangling the testing issues from the political ones is impossible because the consequences of testing are predominantly social and political. However good or bad the measurement of human performance is, it results in socially, politically, and economically important decisions. This endeavor is going to be political. At best we can hope that this turbulent juxtaposition of measurement and scientific concerns and political ones will somehow result in sound policy.

These matters will ultimately be decided in the courts.

DR. LOEWEN: I agree that these decisions are ultimately political. For instance, the male/female verbal score difference is political. Neither cognitive psychology nor testing nor inherent abilities dictates any reason for that difference. In the early seventies, someone at ETS made the political decision that girls should not score higher than boys on the verbal subtest, that it should be the other way. And it has taken several years to find out that the decision was made. ETS admits that that decision was made, but we don't know exactly why. It's better for political decisions to be made politically if that means out in the open with a great deal of contesting.

We're not having a crisis right now in terms of who gets what, when, and why. Admission to colleges is not a big problem. Because of the birth dearth, most colleges accept most applicants. Only about 50 schools are highly competitive, and some very good colleges have empty places.

The symbolic meaning of our policies is the more important issue. When we call the Scholastic Aptitude Test a test of aptitude, we locate the problem in the individual rather than in differences in achievement partly due to unequal schooling, to unequal testing, and to other things. The political struggle is important symbolically.

DR. COLE: First, the statement that in 1972 ETS changed the test contents to disfavor girls is false. That did not occur.[3]

The political dimensions are very complex and difficult. Our society needs to be competitive; our educational system needs to be strong. These goals require high standards in all segments of this society. Because we've fallen down terribly in that, courts will be attacking affirmative action efforts over the next few years.

The issue of scholarship awards to males and females demonstrates the distinction between policy issues and technical issues. Having scholarships awarded to males and females at dramatically different rates is intolerable. However, it does not follow that those tests aren't showing some valid results, especially on math differences, which produce the differences mostly at the extreme levels of scholarship selection. The policy issues require more attention now than test validity.

COMMISSIONER CHAN: Can a fair, universally applicable, unbiased test be designed? Is it possible to construct an idealistic test which will meet the civil rights laws and be fair to everyone?

DR. BOND: I doubt we will ever devise a test that will not disfavor the backgrounds of some children because that is part of what the test should reflect. However, your question really addresses test use rather than the inherent properties of the test. With wise use, biased tests can be used in an unbiased fashion.

Except for admission to a few very selective colleges, the SAT is not as crucial as it once was. It does play a large role in certain scholarship awards. When the SAT is used as a cut score for the awarding of scholarships, I feel it is misused. That's not a wise policy.

DR. COLE: We will not soon determine procedures for allocating the goodies in this society that will satisfy all the people who have a stake in it. Whether tests or anything else play a role, that political issue is a critical part of the way society operates.

The test issue is irrelevant to the fundamental difficulty of the question. But it is ETS policy that tests should not be used single-handedly to represent the diversity of things that ought to be considered in important decisions like this. ETS has spoken out against the improper use and too much reliance being placed on a test score when other information is clearly relevant.

MS. WIGDOR: I'd like to answer your question in its narrowest sense: Will test instruments in the foreseeable future be perfect predictors? No.

COMMISSIONER BUCKLEY: Dr. Cole, you have said that several indicators are more useful than one.

Schools used to have review committees so that if an applicant had a special situation, s/he could apply to the review committee. The committee might review the case and accept the student even with a low SAT score. Why aren't schools still doing this? Why aren't we encouraging schools to look at a multiplicity of things?

DR. COLE: People use single indicators for efficiency. That's true in college admissions. More than 20 State institutions get 20,000 applicants for 4,000 positions. A complex review of applications is not feasible. Admissions offices don't have the time or staff to look at everything that one would want to look at. So people fall back on efficient procedures. More complex procedures are more expensive, more difficult, and take more time, but our pressures ought to be in that direction.

---

[3] At a more recent conference ("Hearing on Gender Bias in Testing," cosponsored by National Women's Law Center and National Commission on Testing and Public Policy, Oct. 13, 1989), Cole commented that those who were then at ETS disagree about why the decision was made. She suggests it may have been to emphasize science, rather than to disfavor girls per se.

COMMISSIONER RAMIREZ: With current levels of test validity and use and the existing adverse impact of tests, can we use tests for disadvantaged, at risk children in grades Pre-K through six, with good conscience, particularly where large numbers of children are being treated by large systems? I'm talking about tests used for ability grouping, used for placement in special education, and used for the acquisition of more funds.

DR. BOND: Suppose we were to get rid of these tests.

At early ages we probably should. Imagine a testless educational system up through the fifth grade. What effect would that have on children's later performance, or on our ability to prescribe individualized educational programs for them? I just don't know. I think some societies are testless and are doing quite well.

DR. COLE: I am uncomfortable with multiple choice testing in grades like K, one and two. To be confident that the schools are teaching children what is important, ask yourself, "Instead of 'the test,' what do I want my children to know?"

Many educational achievement tests are used to make educational decisions about children that are not always optimal educational decisions. That is not the purpose for which achievement tests were designed. Unfortunately, test scores can be used to validate the status quo instead of to change it, which is the purpose of the educational system.

VICE CHAIRMAN FRIEDMAN: In the next session we will discuss legal and policy issues, in contrast to the more technical questions we dealt with earlier.

Presentation of Clint Bolick

MR. BOLICK: We are now able to develop tests that predict job performance and academic performance, and just as rapidly as we are developing them, we abandon them. The abandonment of tests and other objective standards affects our competitiveness as a society, our productivity, our efficiency, our achievement, and equality.

The National Academy of Sciences was asked to address two questions. First, is the General Aptitude Test Battery valid? Second, does race norming the scores detract from that validity? The answers to those questions were fairly obvious before the study was conducted. The first answer was that the GATB correlates positively with achievement in the workplace; and the second, that race norming does detract from the validity.

Yet the National Academy of Sciences created for itself a third question: Does this somehow comport with equality? The National Academy's notion of social equity was not that appearing in the Constitution, which demands equal opportunity, but the notion of some statistical experts. Their answer was no.

The study uses the same strained logic embodied in the Uniform Guidelines of Employee Selection Procedures and in other policy over the last 20 years. It is premised on two notions. First, any statistical disparity is evidence of discrimination, and second, statistical disparity is cured with race conscious relief. The implicit racism and paternalism of these notions ought to trouble us.

In California, *Crawford v. Honig*[4] shows the bad effects of policies founded upon such premises. Black students are being told that, unlike Hispanics, Asians, and whites, they may not take I.Q. tests even when the tests may keep them out of educable mentally retarded (EMR) classes. Our principal client, Demond Crawford, is half Hispanic and half black. He was told if he would reclassify himself as Hispanic, he would be allowed to take the test. We are so far from the notion of equal opportunity embodied in Title VII that we have law cases like *Crawford v. Honig*.

The abandonment of testing, mistaken notions of discrimination and our casual resort to race-conscious remedies leave intact very serious problems in our society. They are the problems of human capital development, of economic mobility and opportunity, and of the abandonment of standards. Human capital development and economic empowerment are the keys to getting people to pass tests. Rather than changing tests to achieve the desired outcome, let's give people the tools, the skills, to pass those tests.

The *Atonio*[5] decision makes legal standards more rational. These standards require a showing of a discriminatory predicate, that is, statistics showing a tendency toward discrimination, before employers must abandon the test or another selection tool. Beyond that, the new legal standards allow the employer to defend the test as nondiscriminatory when he can show that it has a positive correlation with business objectives.

Certainly the General Aptitude Test Battery meets those standards when employers use its scores without attention to race, ethnic group, or gender. With a labor shortage, with businesses willing to invest in training for people who lack skills, the 1990s provide the opportunity to bring people inside the door. However, if we continue to call things discrimination that are not discrimination, to remedy these instances with more discrimination, and to abandon standards, the serious problems in our society will never be resolved.

We should address ourselves to giving people the tools to earn their share of the American dream.

COMMISSIONER GUESS: What is the Landmark Center for Civil Rights that you represent?

MR. BOLICK: The Landmark Center for Civil Rights was founded in May of 1988 to promote equality under law and individual rights.

Thus far we have been challenging barriers to entrepreneurial opportunities that affect those outside the economic mainstream. Earlier this year we successfully challenged a Jim Crow era law here in the District of Columbia. The law prevented individuals from shining shoes on public streets. We are challenging other entrepreneurial barriers, for example, a Texas law that prohibits people from engaging in jitney services. We are looking at cosmetology and other occupations.

We are also involved in a number of cases concerned with equality under law. For example, the *Crawford v. Honig* case challenges California's blacks-only ban on I.Q. tests.

---

[4] No. C-89-0014-RFP (N.D. Cal. 1988).

[5] *Wards Cove Packing v. Atonio*, 490 U.S. 642 (1989); 810 F.2d 1477 (9th Cir. 1987), *cert. granted*, 56 U.S.L.W. 3894 (U.S. June 20, 1988) (No. 87-1387).

The Landmark Center for Civil Rights is supported by private funding—foundations, corporations, individuals, and so forth.

Presentation of Barry L. Goldstein

MR. GOLDSTEIN: In the last 2 weeks four[6] Supreme Court opinions have devastated the protections against employment discrimination that have been available to minorities and women for two decades.

In our competitive society, testing is an important part of how we fairly allocate opportunities, taking into account the civil rights of all concerned and economic productivity. Unfortunately, civil rights groups can no longer concentrate on the technical testing issues. Instead, as a result of the Supreme Court opinions, we are reevaluating principles of fairness and equal opportunity and revisiting issues once thought settled.

I will first put the testing issues within the legal and social contexts, then examine the principles that the Court developed after the Civil Rights Act of 1964 passed, and finally discuss how the recent decisions affect those principles, particularly in the selection area.

I and others in the civil rights field hold three basic premises.

First, our nation's major civil rights problems must be worked out in the courts and not in the streets, and in order to do that we must break down barriers to equal employment opportunity. Congress passed Title VII of the Civil Rights Act of 1964 in this endeavor.

Second, when Title VII was passed, this country's work force was segregated in many jobs. Industries, whether they were power companies, steel companies, foundries, paper manufacturers, or railroads, had jobs that were black jobs and jobs that were white jobs. I'll talk about race, but the same is true of national origin and gender.

Third, various selection practices maintained job segregation. Those selection practices, particularly seniority systems and the use of some tests, were responsible for segregation prior to, and after, 1965.

The Civil Rights Act of 1964 was the response to these three premises. Later, Congress' approach in the Civil Rights Act was mirrored in *Griggs*.[7]

Griggs held that if a plaintiff showed that a selection system disproportionately limits the job opportunities of minorities or women, then the burden shifts to the employer or union using it to justify its use. It doesn't mean, as was previously suggested, that the plaintiff wins, that he has established discrimination, or that an affirmative action plan necessarily follows. It only shifts the burden of justifying the practice to the person who wants to use a system creating those barriers.

---

[6] The four decisions were *Watson v. Fort Worth Bank*, 487 U.S. 977 (1988) (Plurality opinion); *Wards Cove Packing Co. v. Atonio, op.cit.*; *Lorance v. AT&T Technologies, Inc.*, 490 U.S. 900 (1989); and *Patterson v. McLean*, 491 U.S. 164 (1989). Mr. Goldstein commented that the *Lorance* decision makes challenges to intentionally discriminatory seniority systems almost impossible. *Patterson*, he said, permits racial harassment on the job as long as the person was hired and given a nondiscriminatory contract. Although Title VII forbids racial harassment, it's only remedy is lost backpay. So, unless State law provides remedies, all a plaintiff can win is a court injunction for the harassment to stop. *Lorance* and *Patterson* do not bear on testing issues.

[7] *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971).

Griggs makes sense because the employer has the evidence. The employee doesn't have evidence about how or why a selection procedure was used. The employer does. If the employer is only using a selection practice to increase productivity, that is, to select the better workers, and if a test selects better workers, the employer may present that evidence. He wins. That's the *Griggs* rule.

With respect to testing, *Griggs* requires looking at a particular use of the test in a particular job setting for a particular job. For example, a test including questions about Shakespeare is all right for people who are applying to be Shakespeare professors, but very complicated verbal questions are inappropriate for people who are trying to become front-line, blue-collar supervisors. I'm not criticizing all tests or even a particular test.

The *Griggs* ruling had a dramatic effect on the workplace. It changed employers' selection practices, their monitoring of the consequences of those practices, and their rationale for those practices.

It is hard to separate the effects of one particular change in society from others that may also have had an effect. However, some studies have attempted this. Professor Blumrosen compared the work forces in 1980 and 1965. He concluded that nearly a quarter of the minority labor force of 1980 were in significantly better occupations than they would have been under the occupational distribution of 1965.

Jonathan Leonard, a professor of business at the University of Berkeley School of Business, analyzed the effect of Title VII compared to other factors. He concluded that the use of Title VII and the *Griggs* standard in class-action litigation increased the opportunities of minorities in the workplace.

Has the increased minority share in the work force created losses in productivity? If so, our society has problems. However, I have not seen a justification for a loss of productivity. Instead, productivity has increased.

In a 1985 consultation with this Commission, Professor Leonard concluded: "Relative minority and female productivity increased between 1966 and 1977, a period coinciding with Government antidiscrimination policy to increase employment opportunities for members of these groups. There is no significant evidence here to support the contention that this increase in employment equity has had marked efficiency cost."

Another example is a study of the effect of affirmative action on the medical class of 1975, published in the *New England Journal of Medicine* in 1985. It concluded that affirmative action resulted in better health care for minorities because more minority doctors provided care for underserved minority communities. The study also concluded there weren't significant differences in how these minorities did on various tests.

By opening the work force to all segments of the society, we increase productivity. By breaking down barriers to discrimination, we expand the market and get the more qualified people from all groups. We've been underutilizing the productivity of minorities and women both in this society and in helping societies in the Third World. Most of the world would be receptive to America's minorities.

An official of the American Psychological Association, Dr. Goodstein, states, "Psychologists generally agree that the caliber of employment practices in organizations has improved dramatically since publication of the existing Uniform Guidelines in 1978." The guidelines required companies to think about what they did when they adopted selection procedures, not just take things off the shelf, not just follow the advice of untrained managers.

The *Griggs* approach and our law have been followed and cited by courts in England, India, and Israel. Ironically, this aspect of our democratic system is spreading around the world, but we're abandoning it at home.

What is the effect of the Supreme Court's recent decisions? *Atonio* deals with employment selection in a unique industry—salmon fishery and cannery in the wilds of Alaska—far away from any population base. The Court first asked, "What is the appropriate labor market for this industry? Whether minority or nonminority, who are the available, qualified workers?" The Supreme Court's analysis of the appropriate labor market was not objectionable. However, instead of remanding the case back to the lower courts for further findings, the Court wrote law and advisory opinion on principles in fair employment law. This was unnecessary dictum.

The Court addressed three issues, each of which could devastate cases challenging selection practices that limit the opportunities of minorities and women and are not justified by business reasons.

The Court has tampered with the burden of proof when a company's selection procedures operate so that the proportion of minorities selected is much smaller than the proportion of minority applicants. In the past, demonstrating that would be enough to shift the burden to the company. The company would then show that those practices either do not result in limited opportunities for minorities or women, or that they produce better workers and are justified by business reasons. But the Supreme Court now says that showing an adverse impact is not enough.

In the first issue, the plaintiff must now show which of the various selection procedures has limited the opportunities of minorities. Why should the plaintiff have to identify which practice caused the impact when obviously one of them did?

Second, the *Griggs* approach was practical because the employer had the evidence to defend the selection practice. Now, however, the Supreme Court put the burden of proof on the employee. But the company still has the evidence. How can the plaintiff prove the selection practice has no legitimate business reason? Furthermore, if the company destroys the records, the plaintiff can't prove which one of those practices caused the adverse impact on minorities or women.

Third, under *Griggs,* the employer had to show that the practice selected better workers. This was known as "a business necessity" for the use of the practice or "a manifest relationship between the selection practice and the job." Now, the Supreme Court says that the issue is whether a challenged practice serves the legitimate employment goals of the employer in a significant way. It defines this as more than a mere insubstantial justification and less than essential or indispensable. Who knows what that means? When law lacks clarity, people don't do anything. They don't settle cases, they don't change.

Enforcement of the *Griggs* standard will come to a halt for these three reasons in *Atonio.*

In this country, fair employment law is enforced by private lawyers. That's our free enterprise system. If an attorney takes a case and wins, he is paid his fees. Almost every one of the 40 fair employment cases that have gone to the Supreme Court was brought there by a private attorney, not by the Federal Government. (Whether it's a Republican or a Democratic administration doesn't matter.) Will a private attorney take a case in which the law is so skewed against him, in which there's little chance that he will prevail, because he only gets fees if he wins? And there's doubt as to whether he wins. As a result, very few lawyers will be taking impact cases. If somebody comes to me with a problem with a test or a system, I'll be hard put to say that I can help them. They'd better look other places. And that's a shame, because there'll be less scrutiny of selection practices in our country. What that means is that there'll be more intentional discrimination.

Also look around the campuses in this country. There's a growing incidence of overt racial discrimination in the workplace and on our campuses. That will continue. Those are the cases that civil rights lawyers will be limited in handling. Yet, we should seek as much remedy and damages as we can in those cases because litigation will be the only threat to companies to limit the amount of discrimination.

The protections that have existed for 20 years must be restored. We must use the political process to change the law and recover what we thought we had.

Discussion

[Because of a previous engagement, Mr. Bolick departed before this discussion began.]

COMMISSIONER RAMIREZ: Given the current shift in the law, will test validity determine either the further erosion or the reinforcement of whatever protections are left? Does test validity matter anymore, given where the Court is going? Or, is it up to the legislative process?

MR. GOLDSTEIN: The proper use of testing and selection practices is always important for the opportunities of minorities and women.

Now the law cannot reach the improper use of tests, that is, tests that are not job related and that limit opportunities of minorities and women. The chances for minorities and women to challenge improperly used tests are reduced.

Selection practices can be gerrymandered to get any desirable result and have it appear neutral. Once you decide the type of test, how you're going to use the test, and how you'll combine it with other methods, you can pretty much predict the gender and racial composition of your work force in many, many jobs.

COMMISSIONER BUCKLEY: According to projections for the year 2020, 68 percent of the population in the country will be minority and 50 percent of the minority population will be dropping out of high school. A lot of them will do poorly when tested, yet tests are used to hire people, to admit them to universities, and to award high school diplomas. You suggest that disparate impact analysis is no longer effective to say that testing is invalid in hiring practices. Are the employment prospects of minorities almost zero? What will help? How do we prepare?

MR. GOLDSTEIN: This Commission must say what's right with respect to employment and disparate impact analysis. Employers will respond to that. The Commission should encourage employers to closely examine their selection practices and to avoid using one that limits the opportunities of minorities and women unless they have strong justification.

At the turn of the century, this country will have to depend on minority workers. We must figure out how to include people in the work force rather than exclude them. This Commission can emphasize this and that it makes good business sense.

Also, we need a strong Federal fair employment law. Some States have very good civil rights laws that embody the *Griggs* impact analysis. Unfortunately, it is a hodge podge—many States that need those laws, don't have them.

Minority groups and women have to go to Congress. This Supreme Court has overturned effective Supreme Court precedent a half a dozen times in the last 10 years, although never as dramatically as this, and we've gone to Congress and had civil rights bills passed. We can do it again.

COMMISSIONER CHAN: First, I think civil rights laws need interpretations so people can understand them. For example, affirmative action means the employer must find a satisfactory plan for hiring minorities and women. But the law stops there and doesn't give other specific guidelines on how to do it. (It's existed so many years that employers know how to circumvent that particular affirmative action law.) So maybe the Civil Rights Commission should study the interpretation of civil rights laws—their intent and why they were established.

Second, I think every test is a subjective examination by which an organization excludes unneeded personnel. Do we need a third party to put a label on testing material to show that it's unbiased and will conform with civil rights laws when properly used? Could a third party organization, either profit or nonprofit or government organization, certify test materials as unbiased?

MR. GOLDSTEIN: I don't know whether we need a Good Housekeeping Seal of Approval on tests, but we should develop examples of good selection practices, some models that employers and educators could use.

COMMISSIONER DESTRO: Lawyers know the distinction the Court drew between business necessity and reasonableness is important. One is fairly strict and the other is like trying to grab a cloud. You just can't do it.

According to the first panel, courts and testing experts rate a test as good or bad. For example, on a scale from 1 to 10, 3 is not bad. Whether a test is good or bad might not be the right question in employment or education. Isn't the right question, "What are you using the test for?" "How does the test relate to what it is you're trying to show?"

We discussed whether the SAT should have aptitude in its name or not. How do you relate the reasonability test? What mechanism can an enterprising lawyer use given the restraints of time, effort, and cost? How can they attack misuse of tests? Are we locked out because of these tests?

MR. GOLDSTEIN: Some uses of tests may be attacked under this new standard. It will be a very risky undertaking.

It relates to what you suggested about the use of tests and how severe the effect is. For example, judges understand that tests are rather blunt instruments; that often one can use a test to select the unqualified from the qualified. Does somebody have the mathematical facility needed in a technician's job? Does somebody have the ability to read and write needed in a clerical job? You can establish that with some degree of evidence. That would be very hard to challenge under the old standard, and wasn't challenged under the old standard.

If, however, we use that blunt instrument, testing, to find the best technician or clerk out of a thousand applicants, and use rank order to select someone who scores a 98.5 rather than one who scores 98, the adverse impact on minorities is dramatically increased. Some testing experts would defend that use, but a lot would not.

Minorities often pass a test in large numbers at a minimum qualification level. There may be adverse impact at the pass/fail level, but it is worse in the upper range of jobs. Among good jobs where there may be 10 applicants for every 1, 1 percent or less of those passing the test will be minorities. Those in personnel will know that ahead of time. For example, in law school, if you just use the LSATs, and you did it in rank order, what percentage of minorities would you get in your law school? Would it be 1 percent or less? Would that be good for your law school? Would that be good for society? No, and with lawyers, you can argue that paper-and-pencil tests are more relevant than for other jobs in our society. Yet many proponents of testing are trying to use tests to rank order candidates. We know that will just about exclude minorities from government jobs and from lots of other jobs.

Whether we win or lose, civil rights lawyers must try to attack the use of tests like that, when the use is so extreme and the results are so severe. Under the new standards, winning will be a lot harder. Before this decision, we would win. Now, it's up in the air.

COMMISSIONER DESTRO: In my experience with employment litigation and law school admissions, the persons using the tests often don't know what they're using the test for. They're using it as a selection criteria, but if you ask most law school admissions committees what exactly does the LSAT tell you, they'll say, "It's a reasonably accurate predictor of first year grades," but that's it. It doesn't tell you anything more.

MR. GOLDSTEIN: People use written tests with scores for convenience when they get a lot of applicants and want the patina of objectivity. Whether you're selecting people for a police department or a fire department or law school, it's convenient to just go down a list. Convenience and saving some money are not reasons to exclude minorities from police departments or fire departments or law schools.

Another example is the use of physical tests with women. A fire department or a police department will have some physical requirements, but through training women can improve their physical abilities and meet those requirements. However, if test results are rank ordered for running an obstacle course or doing pullups or pulling a firehose, you're not going to get any women, or only a few.

COMMISSIONER CHAN: In California, a commercial company helps you improve your LSAT or SAT. For $450 they almost guarantee a passing grade on the LSAT. How can they do it? Why can't this be done for everyone?

MR. GOLDSTEIN: I just took a course to pass the California Bar. It was a terrific course on how to take that test. They said, "Don't worry about the law, you'll be good enough on the law." The best part was how to answer the questions. They drilled you. I could never have passed that test without taking the course.

In upper middle-class, professional neighborhoods like mine, all the parents and children take these courses to pass tests like the SAT. But the poor kid can't afford $1,000 to take the SAT preparatory course. Yet these courses claim to boost scores 100 points, and I've seen evidence that it's true. It's just unfair.

VICE CHAIRMAN FRIEDMAN: Thank you.

Part III

Papers by Experts

(The views contained in Part III—Papers by Experts—should not be attributed to the United States Commission on Civil Rights, but reflect only the opinions of the authors of the respective papers.)

A Sociological View of Aptitude Tests

By James W. Loewen

This paper discusses issues that college entrance tests raise in our society, as seen by a sociologist who has specialized for a quarter century in race relations and education. Although it buttresses points I made to the United States Commission on Civil Rights on June 16, 1989, the paper can also stand alone.

It is entirely appropriate for the United States Commission on Civil Rights to discuss aptitude testing in our society, but this focus also entails costs. The first sections of my paper describe what is wrong with looking at social and educational inequalities through the lens of aptitude testing. Then I discuss how group differences in aptitude test scores are created. My paper then suggests that creating more "aptitude" in the "low-aptitude" groups is *not* likely to work, not likely to equalize opportunity in our society.

At the center of the civil rights debate in this country at present lies a basic value issue: affirmative action vs. equal opportunity. I will argue that the issues usually raised about aptitude amount to a "soft-shoe routine" that dances around this central value issue without meeting it forthrightly.

The question of test bias also avoids this basic value issue, but test bias is the least defensible aspect of aptitude testing. Of all the impediments that face racial minorities, women, and poor and rural Americans, test bias is the easiest to fix. Hence I will give it considerable attention. Finally, from looking at the causes of test bias, my paper will move to the remedy stage. I will propose three remedies to the problem of unequal adverse impact, none of which requires abandoning aptitude testing.

Do Aptitude Tests Have Adverse Impact?

At the June 16 consultation, Commission members may have noted real convergence between Dr. Nancy Cole, vice president of Educational Testing Service (ETS), and myself, a critic of ETS.

- She agreed with me that the term "aptitude" is a misnomer, hence that the Scholastic Aptitude Test (SAT) needs to be renamed.
- She agreed with me that Differential Item Functioning, performed via "standardization" or the "Mantel-Haenszel statistic," is not a technique to reduce bias.

- I agreed with her that test bias is not the sole or even the most important cause of the adverse impact of tests and test scores on minorities.
- We both agreed, as do all experts whose work I know, that aptitude tests have adverse impact on caste minorities, women, children of poorer families, and rural Americans.[1]

On the November 1987, combined SAT, African Americans scored more than 300 points lower than whites, Native Americans scored about 200 points lower, women scored 57 points lower than men, and Hispanics scored about 125 points lower than Anglos (Rosser, 1989). Rural students also score lower than suburban students; at my university, this difference is about 100 points.

Social scientists disagree as to the causes of these gaps. Some argue that blacks (and perhaps Native Americans, Hispanics, women, and poorer and rural persons) are genetically inferior. Some point to institutional discrimination, from prenatal care through high school libraries. Some allege that deficiencies in family structure and interaction decrease the motivation, verbal agility, or other characteristics of minorities, women, or rural Americans. Others argue that test bias plays an important role. Regardless of the cause, all social scientists agree that aptitude tests show adverse impact.

Some educators have believed that the adverse impact of aptitude testing has diminished, owing to declining admissions pressure on our colleges. It should have. But even in this era of smaller young adult cohorts, aptitude tests still channel students' aspirations and influence their selection. "The truth is that the SATs are the single best predictor of college admissions," according to the recent admissions dean at Princeton University (Wickenden, 1989, p. 153).

This problem of adverse impact is larger than it appears and is going to get worse, for two reasons. First, if the same proportion of caste minorities as whites took the SAT, the disparity between majority and minority scores would be even greater.[2] Second, testimony about the General Aptitude Test Battery (GATB) before the Commission indicated that people of color, women, rural Americans, and poorer Americans may face pencil-and-paper aptitude testing even for jobs like welder or gas station attendant![3]

---

[1] "Caste minority" is Ogbu's (1977) term and refers to African Americans, Native Americans (American Indians), and most Hispanics, particularly Puerto Ricans and Hispanics in the Southwest.

[2] Three ETS researchers use this same reasoning to explain women's lower scores (Burton, Lewis, and Robertson, 1988).

[3] The GATB is said to show small but positive correlations with measures of job performance in hundreds of different working-class jobs. The performance measures are dubious, and the correlations are so small that the amount of performance variance that they are associated with, found by squaring them, is minuscule. (If $r = .3$, then $r^2 = .09$ or just 9%!) Those who still believe such a test has value are invited to ponder this simple question: is a paper-and-pencil test for barbering better evidence than a haircut?

Should These Score Gaps Influence College Admission Rates For Various Groups?

The United States and this Civil Rights Commission must face the issue of adverse impact squarely. It is: *should* access to college education depend upon something correlated closely with race (and with income, gender, and place of residence)?

Between 1969 and 1981, through a crescive and decentralized process, the United States decided it should not. During those years, higher education enormously increased its representation of women, blacks, Hispanics, and Native Americans. This broadening of opportunity was due to changes in white attitudes, not in black (or Hispanic or female) aptitudes. The black movement, woman's movement, American Indian movement, and their corollaries brought about an ideological transformation in the Nation which translated into policy changes within colleges and medical and law schools. If we now allow aptitude test scores to drive admissions policies, then we will see those changes reversed.

How Does Aptitude Testing Mislead Us?

The basic problem a sociologist would note with our use of aptitude testing has to do with its focus of attention. Aptitude tests focus our attention within the oppressed group. More than 50 years ago, Gunnar Myrdal pointed out why this focus does not explain anything about "Negro inferiority":

> Little if anything could be scientifically explained in terms of the peculiarities of the Negroes themselves. . . . All our attempts to reach scientific explanations of why the Negroes are what they are and why they live as they do have regularly led to determinants on the white side of the race line (1944, 1964, p. lxxv).

Today we must still look to white society to understand racial differences in aptitude test scores and the differentials in college-going they can cause.

Testing seems to be an individual act: a student answers an item in the "privacy" of a test site and gets it right or wrong. Thus testing causes us to think individualistically. When we note differences in group means, we think of them as coming from a concatenation of individual responses. This style of thinking leads us to look within the individuals and their "aptitudes" to see what causes their poor (or splendid) scores.

Those scientists who are content with present white (and male, suburban, etc.) dominance in America may think in terms of blaming the individual victim for his/her low aptitude. Unfortunately, the entire framework of aptitude testing also influences those social scientists like myself who think in terms of the social environment and want to change the injustices we see around us. This framework causes us to think in terms of ameliorating the individual victim, so we try to raise his/her aptitude. We suggest girls take more math courses, or provide more nursery schools for inner-city children, or whatever.

Either approach locates the problem within the victim, whether remediable or not, whether due to biology or early childhood environment. Either way, this focus causes us to let our higher education establishment, including aptitude testing, off the hook.

Why Don't Aptitude Score Differences Indicate Group Differences In Aptitudes?

Few sociologists now believe that any major intellectual differences divide people of color from whites, rural people from suburban, poor people from rich, or women from men, whether these differences are ascribed to nature or nurture.[4]

There are many reasons for this sociological doubt. First, some sociologists have had personal experience with minority, rural, poor, or female students, who demonstrate as much aptitude, though not always as much educational achievement, as white males. Second, over the last 50 years, a host of research studies have suggested environmental causes for observed group differences in aptitude. Specific interventions, such as putting an interstate highway through Appalachia, or ensuring that the test givers come from the same group as the test takers, have led to noticeable improvements in aptitude test scores (cf. Whimbey, 1980). So has coaching, a major cause of higher scores much more available to affluent whites (Hammer, 1989). Third, sociologists have become convinced that while aptitude resides within persons, it is a function of societal influences. John Ogbu (1977) has argued cogently that present occupational patterns work backward to influence the next generation. Occupational segregation replicates itself by inculcating the degree of aptitude in the student population that is appropriate to their likely destinations. Fourth, other researchers have shown that even I.Q. is very malleable and can be increased by 30 points by a few months of coaching and higher expectations (Fine, 1975; Whimbey, 1980).

Therefore, when sociologists confront large group differences labelled "aptitude," *of course* we question them. Such differences are not compatible with what we know. We doubt that they are real, that they really show lower aptitude. Many sociological reasons to account for the differences present themselves. Thus we suspect that group differences in "aptitude" scores indicate not differences in aptitude but differences in past opportunities and expectations.

What Ideological Function Do Aptitude Tests Play?

Of course, it may be true that one's social environment, from prenatal care through age 18, has so damaged one as to have caused major differences in aptitude. This amounts to saying, "Well, it's not your fault that you are stupid, but here you are, stupider than we, and there's nothing we can do about it now." Sociologists don't buy this argument either. It's too pat. We suspect its ideological utility. Some time ago, Christopher Jencks put it this way:

> As of 1972, white people still ran the world. Those who have power always prefer to believe that they "deserve" it. . . Some whites apparently feel that if the average white is slightly more adept at certain kinds of abstract reasoning than the average black, this legitimizes the whole structure of white supremacy (1972, p. 83).

---

[4] Scientists do still debate whether genetic differences divide women from men intellectually.

Moreover, as a later section will observe, nonwhites *haven't* proven to be less adept at abstract reasoning. Neither have women. In a society still marked by aspects of racism and sexism, sociologists find it difficult to assess the abstract reasoning or other aptitudes of nonwhites and women without distortion.

Do Aptitude Tests Measure Aptitude?

The only difference between aptitude and achievement tests is this: aptitude tests are more general. College entrance examinations typically test achievement, not aptitude, in general areas of English and math. ACT doesn't call its college entrance examination an "aptitude" test, and Nancy Cole agreed that "aptitude" is a misnomer. Similarly, a student who takes a semester of French (or welding) and is then given a final exam in French (or welding) has taken an achievement test.

In one sense, the French (or welding) test can also be construed as an aptitude test. If two students had the same backgrounds, took the same French (or welding) course, and had the same teacher, and one scored 99, the other 58, then we might justifiably conclude that the first student was more "apt," showed better skills in studying, retaining, and speaking (or coordination, judgment, etc., in welding). If we had to place bets as to which student would be better at learning math, we would doubtless pick the former.[5]

Similarly, if two students come from the same backgrounds, enjoy similar educational preparation, and then take a general English and math test, that test might measure aptitude as well as achievement. That is, the test might measure not only what *has* been mastered, but also the *capability* of mastery shown by each student.

If the two students hail from different backgrounds, then the test measures only achievement, and measures that quite imperfectly.

Sometimes psychologists mistake aptitude and achievement. Michael Cole showed that children (and adults) who have not often heard a word may respond to word association tests with the next word that comes to mind, such as "myriad . . . opportunities." With more common words, they can respond with antonyms or with similar words of the same grammatical class, such as "many . . . few," or "boy . . . girl." Some psychologists think that the latter responses show a "higher" form of reasoning. They assume that people who miss analogy items do so because they use the former kind of reasoning, or because of other reasoning flaws. The first kind of responses need not indicate poor reasoning, however. The mistakes may just be an index of familiarity with the words. Cole went on to note:

> A culture fair test . . . would ensure that the materials used . . . were equivalent in frequency of occurrence for each person being tested. No existing test . . . has ever attempted to tailor its materials to major subcultural groups, let alone individuals. . . . We have long known that asking inner-city children about gazebos and violin-cellos is absurd. But when we see that the same problem arises again in more subtle form with peaches and pears, we begin to seriously doubt the efficacy of ability tests. (1977)

---

[5] Our example does not presuppose that both students have equal motivation, because motivation can be considered part of or basic to "aptitude."

Again, aptitude or ability tests are really measuring background.

How Are Group Differences In Aptitude Created?

Let us follow two children, Frankie and Johnny, from birth to the age at which they might apply to college. Frankie lives in Spanish Harlem, Johnny in Darien, Connecticut. Before they are even born, they are treated unequally. Johnny's mother sees her obstetrician regularly, receives the best current advice ("Stop smoking . . . stay active . . . watch your weight gain . . ."). She follows a diet prescribed for pregnant mothers. Her general health, fitness, and nutrition are good. Frankie's mother gets no medical care and inconsistent advice. Her diet is loaded with sugars and starches. Frankie's mother meets an intern at the hospital and gives birth under anesthesia; Frankie has to be spanked into consciousness. Johnny's mother follows the instructions of her Lamaze group and has "natural" childbirth. Johnny does not come out anesthetized.

The infants return home, to very different homes. Frankie's has lead in the atmosphere, from heavy street traffic; her walls were also painted long ago with lead-based paint. Johnny enjoys his mother's company all day, although he is soon placed in a nearby Waldorf school for a few hours of "enrichment play" each week. Frankie's mother works most days, so she leaves her with a neighbor who "watches children" while watching TV. Frankie's father is not a factor in her life, so she gets no verbal stimulation from him, and her mother is generally too tired for much verbal play when she comes home in the evening.

Frankie and Johnny start school. Johnny's school in Darien recently won a national award for excellence in math instruction. It has a computer for every child. Frankie's school has one computer, used for demonstrations in the library. Some of Frankie's fellow first-graders obviously need more attention than Johnny's peers. Nonetheless, Frankie's first-grade class has almost twice as many students as Johnny's.

Johnny's school system has a rich property tax base, owing to splendid homes and corporate headquarters. Frankie's school is part of a city system that is still struggling financially out of a barely averted bankruptcy a decade ago.

As the children progress through school, Johnny's teachers expect him to know the right answer. They perceive the upper middle-class signals he gives off by his dress, bearing, and "show-and-tell" stories. Frankie's teachers praise her for being attentive, a "good student," but they do not really expect her to be excellent in English, math, or any other academic subject. Each summer, Johnny's parents enroll him in different activities: creative dramatics, computer camp, Outward Bound. Once Frankie goes to Vermont for two weeks as a Fresh Air Child, but otherwise she plays with her friends on the block.

In his junior year, Johnny takes the PSAT and the SAT for the first time. His scores are below average for Darien, totaling just under 1,000, so his father enrolls him in the Princeton Review coaching course. "Of course" Johnny is going to college, hopefully to his father's Ivy League alma mater, "if he can get his scores up."

Now in the fall of their senior years, Frankie and Johnny take the SAT "for real," Frankie for the first time. Her main reason for taking it is that it is required of all students in certain schools that have been placed on "academic probation" by the district board.

Which student will demonstrate greater "aptitude"?

To ask that question is to answer it!

We know that poor prenatal nutrition inhibits intellectual performance (Loehlin, Lindzey, and Spuhler, 1975, pp. 225–26).[6] We know that father-absence hurts SAT scores (Deutsch and Brown, 1967). We know that enriched preschooling causes some of the difference in scores between whites and people of color (*Ibid.*, pp. 304–305). ETS tells us that higher family income is strongly associated with higher SAT scores. We know that summer programs make a difference and help explain why minority test scores drop back farther in the summer (cf. Hayes and Grether, 1969). We know that math teachers subtly challenge boys to work on their own more than they do girls. We know that coaching increases scores, especially Princeton Review coaching, and is less available to minorities (Hammer, 1989). And we haven't even mentioned the differences in test familiarity, motivation to take it, awareness of the test makers' subculture, and dozens of other factors separating Johnny from Frankie—all implied in my little sketch.

It would show remarkable *real* aptitude if Frankie's "aptitude"—her SAT scores—equalled Johnny's!

What About The Aptitude Shown By Asian Americans?

In the 1980s, Asian Americans have done famously well in educational institutions. Their success includes good grades in high school (and earlier), SAT scores approximately equal to whites,[7] and high marks in college. A naive white American view holds that their success proves that America is not "really" racist, that nonwhites can succeed, hence that the problem really lies within blacks (and other caste minorities—Native Americans and Hispanics).

I happened to study a group of Asian Americans in 1967, before their current educational excellence manifested itself. I found that Chinese Americans in Mississippi studied and performed adequately in high school but were not standouts. Then they enrolled in average colleges—Delta State University, Mississippi State University, and the University of Mississippi—where again, they graduated on time, but not with high honors.

One group of Chinese Mississippians stood out, however—children of Chinese American men who had married black women. These "Chinese Negroes," as they were called in those days, were likely to be valedictorians of their (black) high schools. They scored well above black averages on aptitude tests. Then they attended such institutions as Brown and U.C.L.A. in the North or private black colleges like Xavier and Tougaloo in the Deep South.

Expectation is the key to explaining their success. One college student described his earlier (all-black) schooling this way: "The teacher calls on you more often, expecting more from you. So you study harder." The expectation process also operates outside of school, involving people other than teachers, including even oneself. "Over the years, the child tends to meet these expectations, just as many of his darker playmates are meeting expectations that they will be slow or unambitious" (Loewen, 1971, pp. 145, 141).

---

[6] Let me hasten to add that the process is reversible: good nutrition leads to >10 point increases in IQ (Loehlin, Lindzey, and Spuhler, 1975, p. 225).

[7] In November 1987, for example, Asian Americans averaged 936; whites averaged 946. Asians averaged 38 points lower on the verbal, 28 points higher on the math (Rosser, 1989).

I mention this process of social expectation operating in an unusual group of part-Asian children deliberately and strategically. Few will suggest that the particular educational success of Chinese Negroes was genetic. The social nature of the process stands out. That same process helps explain Asian American success in other regions of the country. The opposite process—low expectations by teachers, others, and self—helps explain the low "aptitude" of caste minorities in Mississippi and nationally.

How Does The Aptitude Testing Framework Conceal A Vicious Circle?

At any given point in the vicious circle that depresses the aptitudes of persons other than white males, the social system can seem meritocratic. By limiting our field of vision to the *products* of the social system—the individual aptitudes it measures—aptitude testing reduces our ability to see the broader *causes* of differential aptitudes. Those causes emanate from the social structure. We overlook the ways that inequalities in social structure, from obscenely unequal school finance by social class to subtly different math expectations by gender, cause group differences in aptitude scores.

When we locate these different "aptitudes" within the individuals receiving the scores, it seems appropriate to grant or withhold further favors partly on the basis of the test scores. These favors include national and State scholarships, institutional financial aid grants, and college admission itself. The most important favor is the encouragement to apply to college, or to apply to a "good" college. This encouragement comes from counselors, parents, and peers, based partly on aptitude test scores. Finally the student internalizes this channeling and comes to define him/herself as "not college material" or "not Ivy League material," based partly on test scores (cf. Owen, 1985).

This seems reasonable: some individuals doubtless *are not* college or Ivy League material, although as someone who has taught at Harvard and in rural Mississippi, I doubt that any gulf divides student abilities between those two milieux. Again, when we move from the individual level to the group level, we see how this focus on test scores blinds us to social structural causes. Students think "I'm not good at math," or "I don't test well." Students do *not* conclude, "The social system is biased against me," or "Unequal school finance and lack of role models help cause my V-SAT score to be low." They do not see that aptitude scores are part of a vicious circle that helps perpetuate poverty and educational disadvantage by using the sins of the past to limit opportunities in the future.

Does Aptitude Testing Misdirect Our Attempts At Remedy?

These vicious circles in society do offer an unforeseen benefit: intervention in one area will reverberate through the system to cause improvements elsewhere. Recalling my "Frankie and Johnny" sketch, letting more Puerto Ricans into college will eventually cause more Hispanic children to get better prenatal care, etc., avoiding many of the barriers now confronting Frankie. To take a different example, if a welding test has adverse impact on women and minorities, so it is jettisoned, eventually more girls will see women as welders and more minorities can afford to live in a somewhat better school district.

Again, emphasis on aptitude testing directs our attention away from remedies on the social structural level, such as jettisoning a welding test or changing our method of financing public schooling. Instead, aptitude testing focuses our attention on the individual.

To be sure, individual level remedies are still useful. It is important to improve someone's vocabulary, verbal quickness, and test taking ability, so their verbal aptitude score rises. But this kind of remedy will not and cannot make much improvement in group differences in aptitude test scores, because aptitude tests are norm referenced. ETS constructs and scores the SAT, for example, so its mean is always around 500, its standard deviation about 100. There will *always* be a bottom quartile on a norm-referenced test. Blacks, Hispanics, and Native Americans will be dramatically overrepresented in it for decades. Poor and rural Americans will also be overrepresented. Women will be overrepresented in the bottom quartile on the math section. Therefore, when we use aptitude tests to admit students to higher education, we favor affluent suburban white males, regardless of the motivational level or other skills that some minorities, women, and rural students might bring with them.

Our emphasis on individual remedies perhaps underlies the most ludicrous use of aptitude testing in recent years: the NCAA's Proposition 42. Proposition 42 penalizes high school athletes by denying them athletic scholarships if their SAT scores fall below 700 (out of 1,600) or their ACT scores fall below 15 (out of 36). Effectively, this policy also denies college admission to these students, few of whom are white. (The average for all black students in 1988 was 737.) Incidentally, data from the University of Michigan and elsewhere show that many of these students can do college-level work (Sanoff, 1989).

Among the explanations for the policy is the claim that by denying them aid, Proposition 42 "sends a message" to their high school alma maters, a message that sports is not enough, they must stress academics too. The reasoning is convoluted: after their older siblings get rejected by colleges because of low aptitude test scores, current students will demand better instruction, study harder, and thus improve their scores.

Test scores in many inner-city schools are terrible, to be sure. So are other more important educational outcomes, including high dropout rates and low ability to write effective paragraphs. There are more effective ways to affect these outcomes than by penalizing those few students who have discovered a way to get to college even from such poor educational environments! Again, if our thinking were not beclouded by the individualistic emphasis stemming from aptitude testing, we might redouble our efforts at school desegregation, equal school finance, curricular innovations, and the various institutional approaches that have proven successful in school districts scattered across the Nation.

What Is The Basic Value Clash?

Emphasizing aptitude testing masks the core value issue: the clash between affirmative action and "equal opportunity." I place quotation marks around "equal opportunity" because even though universities are formally equal, even though the United States is formally equal, opportunity remains decidedly unequal for people of color. This is particularly true for the three groups—Native Americans, Hispanics, and African Americans—whose oppression dates to the initial white colonization of the Americas. To a degree, and in subtler ways, opportunity in many fields is also less equal for women.

Thus "equal opportunity" as a policy, meaning the elimination of all formal barriers based on race or sex, really is not equal but maintains unequal opportunity. Because past opportunities have been unequal, aptitude tests "find" greater "aptitude" among affluent white males. Thus aptitude tests allow us to imagine we are manifesting equal opportunity as a society or a college when we are not.

Affirmative action is appropriate and necessary as an antidote to past and ongoing institutional discrimination in our society. Affirmative action goes beyond treating all groups "alike," which we have seen to result in less opportunity for women and persons of non-European descent. Affirmative action means taking steps to counter the existing social structure, with its unequal opportunity. Affirmative action means taking responsibility for the makeup of our institutions, not hiding behind some allegedly scientific or meritocratic test. Affirmative action means admitting a cross section of America (chosen by meritocratic means *within* each group, if we wish). "Equal opportunity" amounts to claiming that aptitude tests are meritocratic and that white (and Asian) males "happen" to show greatest merit!

Equal rights to education and employment should not depend on social science studies. Neither should assertions about "aptitude."

Can We Agree That Test Bias Must Be Eliminated?

Just as a focus on aptitude testing obscures the basic value question, so does a focus on bias in aptitude testing. That is, even persons who disagree that any affirmative action is needed, even those who disagree that adverse impact in testing should be addressed, cannot favor a biased test instrument.

That is one reason why critics of aptitude testing as used today for college admissions, such as myself, emphasize test bias. There is another reason: test bias is perhaps the easiest source of adverse impact to remedy.[8] To remedy the other sources of inequality detailed in my Frankie and Johnny sketch requires everything from major prenatal care programs to massive changes in taxation methods. Test bias, on the other hand, not only *should* but *can* be eliminated relatively easily.

Defining "test bias" can be a complicated task. Jensen's huge tome (785 pages), *Bias In Mental Testing*, doesn't even attempt a definition until page 375. Then he takes care to separate "bias" from "fairness." Jensen then embeds various definitions in a discussion he calls "the most complex in the entire book" (1980, p. 376). In her Commission testimony, Dr. Cole defended ETS by embedding her discussion of bias in a longer treatment of validity. Jensen emphasizes predictive validity and would want ETS to defend its tests, if it could, by showing how they strongly predict college success, however measured. After all, predicting college success is the whole point of aptitude testing for college admission.

---

[8] From the opposite end of the ideological continuum, Arthur Jensen agrees that "biased tests can often be revamped so as to greatly lessen, or even totally eliminate, their bias with respect to a particular subpopulation" (1980, p. ix).

At Columbia University as early as 1901, Clark Wissler and J.M. Cattell correlated aptitude tests with university grades (Hull, 1928). Since then, however, this correlation has been allowed to deteriorate, until it has now become the Achilles heel of the aptitude testing movement. The scant correlation between verbal SAT scores and college grades was noted at least as early as 1937 (Dickter, 1937). At present, the verbal SAT adds *nothing* to prediction, once high school rank and the math SAT score are in the equation. A *good* test *would*!

We must note, however, that to admit on the basis of predictive validity poses an immediate civil rights issue. Predictive validity is not very high: the correlation between first-year college grades (or college graduate rate) and high school grades is .4 or .5; it rises an additional .02 when SAT math scores are added to the equation; adding SAT verbal scores causes no further increase. Squaring the correlation coefficient tells what proportion of the variance in first-year college GPA is associated with these two variables. High school grades and M-SAT scores "explain" $(.42)^2$ to $(.52)^2$ of the overall variance in students' first-year college grades. This is only 16 percent to 25 percent of the variation in first-year grades—and less after that.

Because the correlation is rather low, and because the increase in predictive validity caused by the SAT is minuscule, using predictive validity to determine or define bias amounts to a civil rights problem. At many colleges, African Americans, Native Americans, and Hispanics get worse grades than whites and are more likely to drop out.[9] If we were admissions director at such a school, our ability to predict academic outcomes would increase a bit if we overtly based our predictions on high school grades and race. It would follow that if we overtly barred caste minorities and only admitted whites and Asians, students' GPAs and graduation rates would increase somewhat—perhaps more than the small increase in predictive power resulting from adding SAT scores to high school grades. Most of us would not like the value tradeoff we had thus achieved: a very slight rise in the graduation rate in trade for the overt segregation of the institution. We must realize that when we use the SAT, which is so correlated with race that it functions as an inadvertent measure of affluent Anglo culture, we are inadvertently making precisely the foregoing value tradeoff.

This example shows that test bias must not be defined or studied solely with regard to predictive validity. Such a statistical definition fails to capture much of the common sense meaning that "bias" conveys. I believe a combination of content validity plus the intelligent use of the Golden Rule rule can achieve a reasonably unbiased or balanced aptitude test, however.

Unfortunately, ETS has no way of measuring test bias.

ETS implies that it uses two methods for the "detection and elimination of potentially unfair questions" (ETS, 1987, p. 5): face validity checks and Differential Item Functioning ("DIF").

---

[9] Reasons for their relatively poorer performance may include: racism by professors, culture shock, a "white" curriculum which decreases motivation and intellectual comfort, poorer high school preparation, and financial woes while in college, among others.

However ETS apparently does not review items for content bias.[10] Even if ETS *performed* such a review, it would not suffice, for as Jensen correctly observes, face validity checks are not the proper way to screen items. "Only proper statistical item analysis methods can reliably establish bias" (1980, p. 554).

Is Differential Item Functioning (DIF) A "Proper Statistical Item Analysis Method" To Locate Biased Items?

DIF is not such a method. As Nancy Cole noted in her preliminary paper to the Commission (p. 7), "The DIF analysis itself is not a bias analysis." This is why two ETS researchers recently found that doing DIF made no impact on group means. We should not even be discussing DIF! I will discuss it, only because ETS public relations material presents DIF as if it *were* part of a bias-review or bias-reduction procedure (ETS, 1987).[11]

ETS measures DIF in two ways: "standardization" and the "Mantel-Haenszel statistic."[12] Practically, there is no difference: the two measures correlate almost perfectly ($\underline{r} > .95$). I will discuss "standardization" because it is easier to understand and seems to have become ETS's method of choice. As two ETS researchers put it (Dorans and Kulick, 1983, Abstract), "the primary goal of the standardization approach is to control for differences in subpopulation ability before making comparisons between subpopulation performance on test items." In practice, by "ability" they simply mean "score on the whole test." Thus Dorans and Kulick do not use the female-male difference in performance to examine an item. They "standardize," subtracting the percent correct on the item among boys who scored 200 from the percent correct among girls who scored 200, then the same for boys and girls who scored 210, and so on. Then they sum all differences, weighted by the number of girls in each score category, to calculate d_p the "standardized" difference.

When the two groups have similar overall means, the "standardized" difference between their performance on an item roughly equals the simple difference in percentage correct with which we began. But when the group means differ, then the "standardized" difference usually approximates the original percentage difference on the item *minus the difference in the overall means.*[13]

A problem of terminology afflicts discussions of "standardization." To compare groups matched in ability, age, level of schooling, etc., seems appropriate. Good researchers wouldn't usually compare 6th graders with 12th graders. But overall test score may be a circular measure of "ability." Thus this passage by Dorans and Kulick:

---

[10] See my testimony before the commission. An anonymous reviewer conversant with ETS procedures informed us that ETS does not claim to review items in order to eliminate those that unfairly advantage one group. ETS's "sensitivity review" thus does *not* constitute a content review for bias. See also Scheuneman (n.d.).

[11] It is not clear that ETS has ever removed an item from the SAT owing to bias indicated by DIF. Certainly ETS didn't do so before last year. Also, ETS does not claim to eliminate items pointed to by DIF, without confirmation from face validity analysis.

[12] This use of "standardization" does not mean what statisticians mean by the term. Its meaning also differs completely from that in the glossary of the background paper. Therefore I will place it in quotation marks.

[13] If the difficulty curves differ markedly, then $d_p \neq$ the percentage difference minus the mean difference.

> Standardization with respect to ability level . . . produces a simple total group comparison, like that based on the overall performance column, which is not confounded by differences in group ability. Standardization accomplishes this goal by using the same standard ability distribution for both groups (1983, p. 4).

might better be paraphrased:

"Standardization" by total scores produces a simple group comparison, like that based on the overall performance column, but with the overall group difference removed.

The first passage might lure researchers into imagining that "standardization" is somehow more scientific. This may not be the case.

On occasion, DIF can lead to bizarre results. A study of sex differences on the California Achievement Test provides an example (Green, 1987).[14] Girls did better overall and on individual items. Looking at simple percentage differences, girls outscored boys by ≥5% on 1,233 of the 3,102 different items, while *not one item* favored boys by ≥5%. After "standardization," only 298 of the 3,102 items showed differences greater than 5%, and most of those "favored" boys!

When one group performs dramatically worse than another, such as blacks on the SAT, researchers using DIF are as likely to remove items that favor the lower group as items that particularly hurt them. Accordingly, while DIF is an interesting technique, it is not a tool to locate biased items. DIF removes the adverse impact before looking for adverse impact! There may be no substitute for examining simple percentage differences.

Can DIF Actually Increase Bias?

Far from being a tool for locating or reducing bias, DIF may mask or even increase bias. The reason it may have this effect is simple and statistical.

If an item is added to the verbal SAT that draws on black vocabulary or experience, African Americans might be more likely than white Americans to get that item right. Since the rest of the test contains no items based on black vocabulary, including this item would not make a material change in group means. Therefore DIF analysis would flag this item as biased in favor of African Americans. No one of the many items that draw on peculiarly white vocabulary would stand out under DIF, so they would all remain on the test.

Using the point biserial correlation coefficient to test and knock out items has exactly the same effect. It is more likely to knock out items that favor minorities, women, and poor and rural Americans. As David Owen put it:

---

[14] Green used a different statistical manipulation that had the same effect regarding group means.

> On a multiple-choice test, remember, the correct answer is always right there on the page. If that answer looks right to the wrong people—if low scorers pick it just as often as high scorers—then the question will wash out on the pretest and never make it to a real SAT (1985, p. 124).

Similarly, on the math SAT, using DIF may make it harder to add items on which girls do as well as boys. DIF will knock such items out, because they will look anomalous, compared to "normal" items on which boys do better. Again, examining simple percentage differences may be more useful.

What's Wrong With Looking At Percentage Differences?

Eliminating most of the items with the largest percentage differences will certainly reduce both test bias and adverse impact. This simple technique engenders far more opposition than is justified. Opponents exaggerate its proposed use, to ridicule it. Golden Rule balancing need not degrade the predictive validity or utility of tests. Nor does it leave only easy items. Therefore it will not reduce our capabilities "in a highly competitive international economy" (cf. Rudert, 1989, p. 30)!

Neither I nor any other proponent of percentage differences suggest using them blindly or automatically. For example, I do not suggest that if women or African Americans or Native Americans do badly on algebra items, compared to white males, then we should jettison algebra from math "aptitude" tests! On *some* algebra questions, girls do about as well as boys, while on *other* algebra questions, boys do >10% better. Therefore I would suggest that algebra items on which girls do as well as boys should be selected to replace those on which girls do least well.

Of course, content coverage must be watched. Test makers would not want to decrease coverage of a skill or content area accidentally. But just as ETS altered content coverage in the 1970s to increase male verbal scores, compared to female scores, so ETS could change it back, if necessary, to equalize verbal scores by gender in the 1990s.[15] ETS has a huge bank of test items, with information available as to how different racial, etc., groups have performed on each. ETS could therefore apply the Golden Rule rule and drop those items that have proven to be the worst offenders.

Our research shows that applying the Golden Rule rule to replace items that particularly favor white males can eliminate the gender gap on the SAT-V (Loewen, Rosser, and Katzman, 1988). The same process would reduce the gender gap on the math test by about a third. The black/white gap on the verbal SAT can probably be cut by about 40%, and the math gap by perhaps a third.

---

[15] In our verbal discussion before the Commission, Nancy Cole said, "I can't leave unchallenged the statement that we change[d] the test contents so that the girls would be disfavored back in 1972. I don't care who said that was the case, that did not occur . . . " My statement that ETS *did* make such a change was based on remarks by at least three ETS researchers who were there at the time, which Dr. Cole was not (Dwyer, 1976; Donlon and Angoff, 1971). Dwyer specifically states (1976, p.755), " . . . Sex differences in the verbal section of the SAT, which favored females in the early years, now favor males by a few score points. This change in sex difference parallels development of sex related content specifications." Dr. Cole's unsupported verbal denial does not persuade me to disbelieve these earlier statements by ETS researchers.

Can Output-Based Research Reduce Test Bias and Adverse Impact?

As my testimony suggested, aptitude test makers could do research that correlates items with an output variable—first-year college grades, retention and graduation within 5 years, or overall college GPA. In order to avoid the overt racism that such predictive validity research can entail, as discussed earlier, this output-based research must be done *within*, not across, gender and racial groups.

Dr. Cole suggested this research would take too long. It certainly involves a time lag of 2 years and the collection of freshman grades from colleges. Test makers would also have to contend with different grading standards in different institutions and different fields of study. However, ETS has already collected much relevant data on college performance of previous test takers. ETS has simply never used these data for item analysis.

Item analysis based on output variables must be done intelligently, however, particularly where race is concerned. On many college campuses, as noted earlier, caste minorities earn somewhat worse grades than whites.[16] If these colleges admit students solely on the basis of expected college performance, they will exclude Native Americans, African Americans, and Hispanics, except perhaps for those minority children whose parents are suburban physicians. As I have shown elsewhere (Loewen, 1978), at such colleges it may be impossible to require such high credentials of caste minorities that their graduation rate and college GPAs will equal whites. Thus improving tests by using output-related item analysis will decrease their adverse impact, but will not come close to eliminating it.

Can Mean Balancing Eliminate Adverse Impact?

A third suggestion is much cheaper: mean balancing. Mean balancing would add to the scores of low-scoring groups a constant equal to all or part of the difference between that group's mean and the white male mean. This eliminates adverse impact, from whatever source. The scores as reported back to individuals could include the present score, but the "real" score, to be reported to colleges, would be adjusted to account for the group mean differences.

Mean balancing is functionally equivalent to "within-group scoring," but it conveys a very different symbolic meaning. "Within-group scoring" is exactly how the National Merit Scholarship Corporation (NMSC) has awarded its scholarships for many decades. NMSC employs different cutoff scores (on the PSAT) for each State. If it didn't, it could hardly call itself "national," for it would dole out most of its awards to Connecticut and a few other affluent suburban States. Mississippi and Vermont would watch on the sidelines.

NMSC has achieved geographic diversity, because it has used within-group scoring to equalize conditions between affluent white neighborhoods in Jackson, Mississippi, and Darien, Connecticut. But because NMSC has not employed different cutoffs on racial, sexual, or social class lines, it has shut out rural America, nonwhite America, and female America. Thus NMSC has achieved only one kind of diversity.

---

[16] Footnote 9 suggests reasons.

Like other remedies, mean balancing can be attacked: "You mean, just *give* Native Americans 200 points?! Without them *earning* it?!" No one now attacks the National Merit Scholarship Corporation for geographic balancing, however. Mean balancing also carries some advantages. It decreases the stigma and self-fulfilling prophecies that accompany poor test scores. Today an "average" black female scores 371 on the verbal SAT and thinks she is not college material, certainly not "good" college material. With mean balancing, she would score 457. Her 371 would tell her the low percentile rank she fell into, in verbal achievement, so she would know that she had work to do to catch up to the national mean. Her 457 would tell her that her verbal aptitude was in the 50th percentile once her score was adjusted. Moreover, reporting both scores would make Americans of all racial and gender identities more aware of the influence of social structure on the individual.

We have seen that mean balancing is functionally equivalent to "within-group scoring." In my experience, within groups, the SAT does a reasonable job of putting people in rank order, confirmed by their course work. Across groups, the SAT fails. Thus using aptitude tests with mean balancing would maintain our present emphasis on meritocracy. Although women would get a boost in their math scores, they would still compete for positions against other women and against men. Although African Americans would get a sum added to both of their scores, this amount would merely equalize the playing field. The outstanding African American would now score about the same as the outstanding Caucasian.[17]

Some Remedy Is Urgently Required!

As a sociologist, I cannot ignore the prognostic uses of aptitude tests. Proponents of aptitude tests see them as merely the messenger of bad tidings. It would be wonderful if we took them seriously as messengers. Then, although this reform lies far beyond the power of testing agencies, the Federal Government and the States could use group differences in aptitude test results to funnel money, people, and ideas into school districts, schools, and neighborhoods whose scores indicated greater need.

Unfortunately, aptitude tests are much more harmful than mere messengers. Sociologists don't want observed "aptitude" gaps to prescribe inequality for the *next* generation. When we use test scores prognostically—when we admit some persons and reject others—we inadvertently do just that. Using test scores without any of the remedies I have proposed is guaranteed to maintain adverse impact. It's not fair, it's not a good talent search, and it's not good policy for our nation as we strive to hold together as a country.

There are reasons to abandon aptitude testing altogether (cf Crouse and Trusheim, 1988). There are also reasons to retain it. I have not taken a stand on that complex question. But some remedy to its adverse impact on minorities and women is urgently required! The remedies I have proposed:

---

[17] Mean balancing could also make smaller adjustments than the mean differences, if it was deemed appropriate to make up only part of the mean differences.

- reducing the adverse impact of tests by dropping those items with the most adverse impact (Golden Rule rule),
- improving the tests by doing item analysis with output variables, and
- mean balancing,

are not mutually exclusive. ETS (and ACT) could institute all three at once. If none of them are put into place, by law or voluntary action, then as a sociologist I would have to recommend that aptitude testing be abandoned, at least for higher education admissions.

References

Burton, N.W., C. Lewis, and N. Robertson. 1988. "Sex Differences in SAT Scores." NY: CEEB.

Cole, Michael. 1977. "Culture, Cognition, and I.Q. Testing," pp. 116–23 of *The Myth of Measurability.* NY: Hart Publ., 1977.

Crouse and Trusheim. 1988. *The Case Against the SAT.* Chicago: University of Chicago Press.

Deutsch, Martin, and Bert Brown. 1967. Social Influences in Negro-White Intelligence Differences. Pp. 295–307 of Martin Deutsch, *et al., The Disadvantaged Child.* NY: Basic Books.

Dickter, M.R. 1937. *The Relationship Between Scores on the SAT and Marks in Math and Science.* Philadelphia: University of Pennsylvania Ph.D. Dissertation.

Donlon, T.F., and W. H. Angoff. 1971. The Scholastic Aptitude Test. Pp. 15–47 of Angoff, ed., *The College Board Admissions Testing Program.* NY: CEEB.

Dwyer, Carol. 1976. Test Content and Sex Differences in Reading. *The Reading Teacher,* May, pp. 753–57.

ETS. 1987. *Developing a Test.* Princeton: Educational Testing Service.

Fine, Benjamin. 1975. *The Stranglehold of the I.Q.* Garden City, NY: Doubleday.

Green, D.R. 1987. Sex Differences in Item Performance on a Standardized Achievement Battery. New York: paper presented at the annual meeting of the American Psychological Association.

Hammer, Joshua. 1989. Cram Scam. *New Republic,* 4/24/89, pp. 15–18.

Hayes, D., and L. Grether, 1969. The School Year And Vacation: When Do Students Learn? NY: ESS, cited in Ashley Montague, ed., *Race and I.Q.* NY: Oxford University Press, 1975.

Hu, P., and N. Dorans. 1989. The Effect of Deleting Items with Extreme Differential Item Functioning on Equating Functions and Reported Score Distributions. San Francisco: paper presented at annual meeting of the American Educational Research Association.

Hull, Clark L. 1928. *Aptitude Testing.* Yonkers: World Book Co. Jencks, Christopher, *et al.* 1972. *Inequality.* NY: Basic Books.

Jensen, Arthur. 1980. *Bias In Mental Testing.* NY: Free Press.

Loehlin, John, Gardner Lindzey, and J.N. Spuhler. 1975. *Race Differences in Intelligence.* San Francisco: W.H. Freeman.

Loewen, James. 1971, 1987. *The Mississippi Chinese: Between Black and White.* Cambridge: Harvard University Press; Prospect Heights, IL: Waveland Press.

Loewen, James. 1978. Breaking the Vicious Circle. *Clearinghouse for Civil Rights Research, v.6* no. 1–2, pp. 24–35.

Loewen, James, Phyllis Rosser, and John Katzman. 1988. Gender bias in SAT Items. New Orleans: paper presented at the annual meeting of the American Educational Research Association.

Myrdal, Gunnar. 1944, 1964. *An American Dilemma.* NY: McGraw-Hill.

Ogbu, John. 1977. *Minority Education and Caste.* Orlando: Academic Press.

Owen, David. 1985. *None of the Above: Behind the Myth of Scholastic Aptitude.* Boston: Houghton Mifflin.

Rosser, Phyllis. 1989. *The SAT Gender Gap: Identifying the Causes.* Washington, DC: Center for Women Policy Studies.

Rudert, Eileen. 1989. The Validity of Testing in Education and Employment. Washington, DC: United States Commission on Civil Rights, Background Paper for Consultation.

Sanoff, Alvin. 1989. When is the playing field too level? *U.S. News and World Report,* 1/30/89, pp. 68–69.

Scheuneman, Janice. n.d. A Systematic Procedure Aimed Toward Sex Fair Testing. Princeton?: ETS? (draft photocopy).

Whimbey, Arthur. 1980. *Intelligence Can Be Taught.* NY: Dutton.

Wickenden, James W. 1989. Breaking the myths of Admissions. *Money,* 5/89, pp. 153–55.

Judging Test Use for Fairness

By Nancy S. Cole[*]
Educational Testing Service

The basic question before the United States Commission on Civil Rights in judging the fairness of tests is: What does it mean to say either that a use of a test is valid or that it is biased? To provide information to help answer that question, this paper will address five major topics:

1. The meaning of the words "valid" and "biased."
2. The types of information needed to infer that a test use is either valid or biased.
3. The reasons group differences in scores are *not* necessarily indicators of bias.
4. Appropriate ways to judge bias.
5. What we should do about group differences in test scores.

The paper will demonstrate that validity and fairness are inherently linked (as are invalidity and bias), and that judgments concerning a test's validity and fairness should depend directly on the types of inferences to be made on the basis of the scores. The evidence for fairness or bias that should be considered extends far beyond the existence of score differences between groups and includes information about the context surrounding test use, the content of the test, the way the test is administered and scored, the relationships among parts of the test, and the relationships of test scores to external criteria such as grades in college or performance on the job. Differences between groups in test scores provide important information that should not be covered-up by automatically blaming bias for the differences. There are major inequities in education that lead to group differences in test performance. The information derived from tests brings those inequities to our attention and provides information to help us address them effectively.

What Do We Mean by "Validity" and "Bias"?

Our dictionaries tell us that the word valid means "well-grounded or justifiable . . . correctly derived from premises . . . sound, cogent, convincing, telling" (Webster's New Collegiate Dictionary, 1979). By contrast, bias is described there as "a highly personal and unreasoned distortion of judgment: prejudice." The uses of these two terms in relation to tests carry the same connotations: A valid test gives results that are justifiable and sound; a biased test gives results that are unfairly distorted.

However, even given these straightforward definitions, validity and bias are not qualities that we can automatically recognize. To do so we must have clear understanding of what it means to be "sound" and what it means to be "unfairly distorted" in the particular situation in which the test is being used.

---

[*] The author is indebted to Michael J. Zieky for many helpful suggestions in the preparation of this paper.

Cole and Moss (1989) illustrated the complexity of the problem in the following example:

> Suppose one group of high school students, Group A, scored higher on a high school achievement test than another group, Group B. Such an event might lead to a headline in the local newspaper, "Group B Students Score Lower." Callers on a local radio talk show might say "I always knew those Group B students were dumber," or "The schools are not doing a good job with those Group B students." A letter to the editor in the local newspaper might argue that "the test score results do not mean anything because those tests are biased." (p. 201)

We can imagine situations in which we would expect that the test score differences were valid and other situations in which we would expect that the score differences were caused by bias. One situation that would produce an expectation of validity, for example, would be if Group A consisted of students with an "A" grade average in high school and Group B consisted of students with a "B" average. We expect that "A" students would have learned more than "B" students and would therefore score higher on a sound test reflecting that learning. In fact, if there were *no* difference in such a situation, it would raise a question about the soundness or validity of the achievement test or the grading or both.

On the other hand, if Group A consisted of right-handed students and Group B consisted of left-handed students, we might well question the validity of the test score differences. Although it might be possible that right-handed students do achieve more than left-handed students, before accepting that conclusion we would want to be sure that the test was valid and fair. For example, we might wonder if right-handers actually made better high school grades. If right- and left-handed students had comparable school grades, we would have even more questions about the validity of the test. Then we would surely explore in detail the possibility of bias in the test or its administration. For example, we might wonder whether the students took the test in right-handed chair-desks and whether the chairs unfairly handicapped the left-handers.

Judging when a use of a test is valid (justifiable, sound) and when it is biased (unfair) is a difficult and complex process. It requires a clear understanding of what we mean by "valid" and "biased," a variety of types of information about the test and the situation in which it is used, and a recognition that score differences, in and of themselves, are neither an indication of bias nor of validity. Let's look then in more detail at what we mean by "valid" and "biased" in testing and how that can assist us in trying to ferret out bias in practice.

According to the 1985 *Standards for Educational and Psychological Testing,* validity:

> refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences made from the scores. The inferences regarding specific uses of a test are validated, not the test itself (American Educational Research Association et al., 1985, p. 9).

This definition makes several important points that deserve emphasis.

Valid for What? Validity is not a characteristic of a test but of inferences based on the test scores. Thus, it is not the test itself that is found valid or not valid but specific inferences from the test scores. The statement "A test is valid or invalid" is not appropriate without the description of the inferences for which it is valid or invalid.

The focus on "inferences" requires that we identify the inferences being made, subject them to scrutiny, and search for evidence of their appropriateness. We should be alert to implicit, unexamined inferences that may be made on the basis of scores.

Types of Inferences. The fact that different types of inferences are made on the basis of test scores is one of the complicating factors in defining what we mean by validity. At a first level, inferences refer to the immediate meaning given the score. Typically, this meaning is in the form of a person's level on some characteristic such as "math skill" for a math achievement test, "intelligence" for an IQ test, or "assertiveness" for a personality test.

At a second level, inferences reach beyond the immediate test score meaning and present state of the individual to some further-removed inference such as whether an educational intervention is likely to work for the individual. This level of inference involves not only what the test score is supposed to mean immediately, but how that score interacts with external factors. The logic of validation makes clear that we should examine evidence to determine whether this second-level inference is correct or not. Note, however, that a second-level inference might be invalid either because the test is not working as planned or because the intervention is not working as planned. In either case the inference is incorrect. However, in one case we try to change the test; in the other, we try to change the intervention.

A third level of inference or expectation involves more distant expectations with respect to some ultimate purpose. For example, teachers make inferences about children in classrooms each day and take action based on those inferences. They also have expectations that their actions will result in certain long-term benefits such as making the students become productive adults. However, we rarely subject the longer term expectations about the overall educational (or social) good of such inferences and actions to the validity requirements of evidence. The same is true with testing practices. Most educators have more distant expectations with respect to the educational or social good of particular testing practices, but such expectations are rarely validated.

Taken together, the three levels of inference suggest the wide range of considerations in test use including concern with unintended as well as intended outcomes and examination of evidence about outcomes at different levels of inference.

Multiple Sources of Evidence. Various forms of evidence are relevant to judging the appropriateness of an inference. Different forms of evidence are needed to address the various types and levels of inference noted above and to address the very different contexts in which the test might be used. The various types of inference relate to the distinction between content validity evidence (for inferences based on the content of a test such as algebra problems), construct validity evidence (for inferences about traits such as quantitative ability), and criterion-related validity evidence (for inferences about predicting a criterion such as grades in college).

Although in the past, users sometimes thought it sufficient to select one type of evidence and examine it alone, the *Standards* make clear that validation is a "unitary concept" referring to the evaluation of all the evidence about an inference. In addition, the field is increasingly recognizing the need to include a thorough examination of the context of the use including the characteristics of the examiners and examinees and the situation in which the test use occurs. Cole and Moss (1989) referred to this as the need for a "context-based" unified validation.

Implications for Definition of Bias. To be valid, a test must be fair.[1] Validity refers to the appropriateness of an inference from a test score. To be appropriate, an inference must be unbiased or fair. Fairness is a necessary condition for validity. In the testing context, concerns of fairness are concerns about the appropriateness of test score inferences for particular groups. For example, we may wonder if the test is fair for individuals of different racial or ethnic identity, for persons of different levels of economic advantage, for persons of both genders, or for persons with particular physical handicaps.

Validity concerns the appropriateness of inferences about examinees in general. Fairness, as a special subset of validity, concerns the appropriateness of inferences for special groups of examinees. Bias, then, is a particular type of invalidity—invalidity or differential validity with respect to particular groups of concern.

Fairness is the logical counterpart of validity (as bias is of invalidity) and the same issues drive the examination of both validity and fairness. Consequently, it will be convenient to speak of issues of validity and fairness in tandem in the following section in which the information needed to examine both is addressed.

What Information about Validity and Fairness is Needed?

Consider again the introductory example about test score differences for Group A and Group B students. One immediate inference is that Group A students have learned more in the subject of the achievement test than have Group B students. The validity (and fairness) issue is: Is that an appropriate inference? It was clear in the example that we needed to know a good bit about the situation, including who the students are in Group A and B and how they otherwise perform academically. When Group A and B were students who had "A" and "B" averages, respectively, in high school, we approached the issue rather differently than when we considered left-handed and right-handed students. The latter situation raised special issues about the conditions under which the test was given.

This illustrates the many types of information that need to be considered in judging validity (and fairness). Validity and fairness (or invalidity and bias) cannot be represented by a single number from a single approach. The questions we must ask do not result in simple yes or no answers. They are complex and involve many types of information that must be consolidated by knowledgeable judgment into an overall decision about whether there is sufficient evidence to support an inference from a test score.

A major purpose of this section is to indicate the areas of evidence related to validity and fairness that should be examined. Sometimes a single piece of information is wrongly treated as the sole answer to validity and fairness issues. This section also provides the context and perspective from which single evidential pieces need to be judged in the overall judgment of validity and fairness.

---

[1] Although some writers have differentiated the concerns of bias and those of fairness, in this paper the words "fair" and "unbiased" are used as synonyms as are the words "biased" and "unfair."

In a recent chapter on "Bias in Test Use," Cole and Moss (1989) identified five general areas in which we would need various types of information and evidence to judge validity and fairness. These five areas provide a guiding framework for evaluating tests.

The Context of Test Use. The first concern is to understand the context of the use sufficiently to know what questions need to be asked about validity. An interpretation of a test score takes on various shades of meaning because of the context in which the interpretation is made. Uses of test scores for self-evaluation raise different questions than do uses for selection. Within each use category, the particular use raises different questions. For example, we would have a different set of questions to ask about the use of a test for selecting secretaries than for the use of a test for selecting unskilled laborers. To ask the right validity questions, we must understand the context—with whom the test is to be used, under what conditions and for what purpose, what action will be taken on the basis of the scores, etc.

Content and Format. The second area concerns the appropriateness of the test content and the formats of the questions for the particular interpretation of the scores to be made. Here, for example, we explore the types of math included on a math test and the form of the test questions used. This category includes the area referred to as "content validity," expert judgment about the content of the questions. It also includes evidence about the appropriateness of the content for various groups of concern (the fairness issue).

Good test development procedures provide multiple examinations of content. Panels of experts in the subject matter define the appropriate content for an achievement test, for example. Such panels must provide sufficient breadth to represent the variations in what is taught from school to school for a statewide test, for example, or from classroom to classroom for a local test. Fairness issues involve possible differences in the content accessible to different groups of concern.

In addition, good practice includes having content subjected to a special review for possibly offensive content, for possible stereotyping of groups, for balanced references to different groups where appropriate, and other such content concerns related to fairness issues. At Educational Testing Service, for example, this type of review is called a "sensitivity" review. Special reviewers receive training in the issues of a sensitivity review and all tests are subjected to such a review by trained reviewers independent of the persons responsible for assembling the test.

Another area included in this content and format category involves effects on test takers of the contexts in which tasks are set or the way questions are asked. These considerations focus on whether there might be special characteristics of the question context or format that differentially affect people in different groups. For example, whether or not mathematics questions set in sports-related contexts are equally appropriate for females and males is the type of issue addressed here. Similarly, the question of group differences related to different forms of questions (multiple-choice versus essay) would be another set of issues for this category.

Administration and Scoring. The third area involves the way a test is given and scored. These, too, are important factors in what a resulting score means and whether it shows forms of bias. The basic concern of standardization is that the test be given and scored so that all test takers are treated the same way. Standardization is therefore a basic fairness concern. In addition, it is critical that procedures be consistent with the intended inferences to be made from the scores. Many types of information are sought to check that the procedures produce the intended meaning and are comparable and fair to all examinees. For example, there are studies of the effects of the racial/ethnic identity of the test giver on the performance of test takers of the same or different identity. The previous example of the right-handed/left-handed groups raised issues of the testing situation (namely, the nature of the chairs) that might produce differential results. The amount of time allowed examinees fits this general category, too. Currently, issues of the possible differential effects of time limits on tests that require students to work quickly are the target of attention and study.

Internal Test Structure. If a test and questions on a test are intended to have a particular meaning, then that meaning implies certain relationships among test parts. For example, a math test might include a total score as well as subscores on problem solving and computation. A question that is part of the problem-solving section should be more highly related to the problem-solving score than to the computation score. As part of the total score, each question should be related to it as well. These are the types of issues addressed in internal test structure.

A variety of statistical analysis procedures are used to investigate the properties of individual questions or clusters of questions in relation to the test as a whole. The purpose is to see if the intended and expected relationships and properties exist.

Many widely discussed methods to examine possible "item bias" (unfair questions) fall into this category of information as well. They involve how responses by groups to particular questions relate to other internal characteristics of the test. When internal relationships are similar for different groups, fairness is supported. When such relationships differ by group, questions of bias are raised.

External Test Relationships. An important part of the information about a test's validity and fairness (or invalidity and bias) for a particular use concerns how test scores are related to measures external to the test. For example, if a test is supposed to measure preparation for college work, then how the scores relate to eventual college performance is an important issue. This relationship is typically labeled criterion-related or predictive validity. Relationship of the test scores to performance in high school or other such external variables may also help explain what the test is measuring and how it should be interpreted.

Fairness issues involve the relationship of test scores with external variables for special groups of concern such as women or Asian examinees. For example, there have been many studies of the prediction of college grades by test scores for different groups of concern. This whole line of research illustrates the types of information and evidence appropriate to the external relationships category.

Summary. In summary, many types of information must be considered to address issues of validity and fairness. We should not expect to find an answer to validity or fairness issues in only one type of information or one single number. We must judge a variety of information to reach an overall judgment about whether there is sufficient evidence of validity and fairness to support a particular test interpretation or inference.

Why are Group Differences on Tests Not Necessarily a Sign of Bias?

In spite of recognition by measurement experts of the wide range of information that should be considered to judge validity and fairness, much public attention has focused on differential performance by groups on test scores or on individual test questions. The major message of this section is: *Raw differences between groups on test score averages or on individual test questions are inconclusive by themselves for judging the fairness of the test or the question.* Such differences are often important for other reasons, but they help little, if at all, with issues of bias.

In certain examples, it is easy to see that we should not automatically conclude that a test score is biased for an inference just because groups differ on it. Suppose a test measured the heights of males and females. When the results showed that males tended to be taller, would we conclude that the test was biased against females for the inference about relative height? Certainly not. However, if the inference being made from the height scores were ability to do a particular job in which height played no role, that use of the test would be invalid and unfair.

Or suppose there were differences in mathematics test scores between tenth graders and seventh graders. Would that mean the math test was biased against seventh graders in the inference that they knew less mathematics? No. However, if the inference were that the tenth graders were better able to learn mathematics, serious questions would be raised. As another example, we would not require that to be a valid and fair measure of Spanish fluency, a test would have to produce identical scores for native English speakers as for native speakers of Spanish.

To require equivalences between groups with respect to an inference without regard to possible valid score differences is unreasonable as these examples show. It is wrong to assume that all groups will score the same on every test—even though our social concerns might lead us to wish this were the case. On the other hand, as each of these examples show, it depends on the particular inference being made as to whether the group differences provide valid or invalid inferences. Thus, we have to look very closely at the particular inferences being made along with the group score differences to examine reasonably the validity and fairness issue.

How Should We Judge the Validity and Fairness of Inferences?

Of course, the examples given above are relatively free of complicating factors. None of these are the issues before the public today with respect to the possible unfairness of inferences from test scores. Let us consider two of the more difficult cases of group differences that do represent some of today's primary concerns:

Case 1: Different performance on academic tests by members of racial-ethnic minorities.
Case 2: Different performance on academic tests by females and males.

For each case we examine the types of information relevant to judging validity and fairness and the conclusions that can be reached in specific instances in which considerable evidence about validity and fairness is available.

Different Performance on Academic Tests by Members of Racial-Ethnic Minorities. There have been concerns for decades about the relatively lower test scores of black students compared to white students and of students from less advantaged social and economic family conditions compared to students from more advantaged social and economic family conditions. Such concerns have been fed by score differences on a range of tests from so-called intelligence or IQ tests to tests of achievement (what has been learned to date) in school-related subjects.

The types of tests illustrate the different inferences possible to draw from test scores. Such different inferences are the basis of much of the public concern with respect to possible bias. For example, the inference that test score differences represent different abilities to learn or different levels of intelligence created a furor in the late sixties and early seventies.[2] The obvious concern was that such inferences would lead to teachers and schools "giving up" on members of lower scoring groups on the theory that they were not able to learn anyway.

We have come a long way since those early discussions. There is a broader understanding that any test is directly measuring only what a student can do at a particular time. Whenever opportunities to learn differ between groups, score differences will reflect those different opportunities. Thus, educational tests focus on what students can now do without the implication of what they could or could not have done under different circumstances.

The Scholastic Aptitude Test (SAT) is a case in point that receives considerable attention. The focus of the SAT is on developed general verbal and mathematical reasoning abilities important to college work. To some, the term "aptitude" wrongly seemed a synonym for "intelligence." More appropriately, it represents a focus on a prediction of college performance. As used and interpreted today, the SAT might just as well stand for Scholastic Achievement Test—it measures the general achievement of students in verbal and quantitative reasoning rather than specific achievement in a particular subject, but achievement nonetheless.

When SAT score differences are found between blacks and whites, the first question is what inference is being made. One possible inference is that black students and white students, if given the same previous educational opportunities, would have different prospects of success in college. Since there is abundant evidence that, in general, black and white students do not have the same educational opportunities, it is clear that we cannot assume the same opportunities in such comparisons. Such an inference goes well beyond present supporting evidence and is an inappropriate and biased inference.

However, another possible inference is that black students are, on average, not as well prepared for college work today as are white students on average. For this inference to be validated, we would need to look at the range of types of validity information including the context of test use, content and format of the questions, administration and scoring of the test, internal test structure, and external test relationships. For illustrative purposes here, let us consider partial evidence from two of these categories of information: measures of "item bias" and predictions of performance in college.

---

[2] Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–123.

There is no statistic that can prove whether or not a test question is biased. Simple differences between groups in the percentages of examinees answering a question correctly can not be used as proof of bias because the groups may really differ in knowledge of what the question is measuring. Efforts to use such simple differences as indications of bias have been rejected by professionals in the field of measurement for precisely that reason. If, however, examinees could be matched in terms of relevant knowledge and skill, then people in the *matched* groups could generally be expected to perform in similar ways on individual test questions. Methods described in the technical literature for identifying questions that may be biased generally use some form of group matching before calculating differences in the difficulties of questions between groups.[3]

What is called "differential item functioning" (DIF) occurs when people of approximately equal knowledge and skill (matched on relevant factors) in different groups perform in substantially different ways on a test question. Measures of DIF thus help to identify questions that may be biased because group differences in relevant knowledge and skill have been taken into account to the extent allowed by the matching process.

When new versions of the SAT are assembled, test developers use DIF results so they can avoid use of questions with high DIF values for some evaluated group. The use of such procedures cannot, of course, guarantee that the SAT will be a fair test. The use of such procedures does, however, add to the mix of evidence that can be gathered to demonstrate the fairness of the inferences made on the basis of SAT scores.

The SAT is designed to allow inferences about the future performance of high school students in college. Crucial evidence for the validity and fairness of the test must come from the relationships between SAT scores and first-year grades in college for people in various groups. Such evidence was summarized in a report of the Committee on Ability Testing under the auspices of the National Research Council:

> The observed differences in score distributions between various subpopulations raises questions of validity and questions of fairness. Whether test data are appropriately used in admissions decisions regarding minority applicants is first of all a factual question: Are predictions made from test scores as accurate for minority as for majority applicants? On the basis of the evidence currently available, the answer is yes. . . . That evidence dispels two contentions regarding within-group and between-group comparisons.
>
> One contention, which pertains to within-group validities, is that tests do not predict which of the black students will achieve the best college records. In fact, however, predictions for blacks as a group are as accurate as predictions for whites as a group. Hence, insofar as admissions officials want predictive

---

[3] For a general discussion of those issues, see Shepard, L.A. "Definitions of bias." In R.A. Berk (ed.), *Handbook of methods for detecting test bias.* Baltimore: Johns Hopkins University Press, 1982, pp. 9–30. For a survey of the range of technical approaches to comparing items across groups (all of which use some procedure to match groups), see Cole, N.S. and Moss, P.A. "Bias in test use." In R.L. Linn (ed.), *Educational Measurement.* Washington, D.C.: American Council on Education, 1989, pp. 201–19.

information to improve the comparison of competing applicants from the same ethnic group, tests provide useful data.

The second contention, which pertains to between-group comparisons, is that experience tables based on the general student population understate the probable success of black students. However, the bulk of the evidence concerning commonly used admissions tests suggests that their predictive validity differs at most only very slightly for blacks and whites. With the important qualification that only scanty evidence is available for minorities other than blacks, subgroup differences in average ability test scores seem to predict similar differences in academic performance as measured by course grades.[4]

Different Performance on Academic Tests by Females and Males. Even though both male and female students receive the full range of scores on tests of all types, average differences in their scores are found. The major score differences are in quantitative areas such as mathematics and science in which male students tend to outscore female students. In verbal areas such as reading and writing, female students tend to outscore male students. These differences appear as early as elementary school and some differences (e.g., math) widen as the students mature.

The differences are found on many tests at different levels, but there has recently been a great deal of interest in gender differences on the SAT. Currently, the male average score in mathematics is between 40 and 50 points higher than the female average, and the male verbal average is about 10 points higher than the female average. (Those differences are on a scale that spans 600 points and would correspond to differences of about 7 to 8 points in math and about 1 or 2 points in verbal on a more familiar 100 point scale.)

As noted, the SAT is designed to allow inferences about examinees' future performance in college. Is the SAT biased against women in making those inferences? As we have seen, evidence from many sources has to be evaluated.

Evidence from the context of test use includes information about the people who take the test. It is important to remember that the SAT is *not* taken by a representative sample of people. Male and female students decide whether or not to take the test.

The young men and women who decide to take the SAT are not representative of all young men and women, nor are the men test takers and women test takers comparable to each other.

The men who take the SAT, for example, are more likely to take high school courses in trigonometry, precalculus, calculus, and computer mathematics than are the women who take the SAT. Given the differences in courses taken, the average difference in mathematics scores may reflect real differences in preparation rather than gender bias in the test.

The men and women who choose to take the SAT differ in other ways as well. More women than men take the test and, as compared with the men, the women are less likely to have attended private schools, less likely to have college-educated parents, less likely to be members of the majority racial group, and less likely to be members of relatively affluent households.

---

[4] Wigdor, A., and Garner, W. (eds.) *Ability Testing: Uses, Consequences, and Controversies.* Washington, DC: National Academy Press, 1982, pp. 195–96.

Clearly, the differences between the men and women who take the SAT provide evidence to help judge the fairness of the differences in their scores.

With regard to the content of the test, the DIF analyses described above are completed for male-female differences as well as for black-white differences. Test questions that show elevated values of the DIF statistic are not used in assembling new editions of the SAT. No statistic can guarantee that there is no gender bias in the questions. The use of the statistic does, however, add to the evidence that can be gathered concerning the fairness of the test and of the questions in it.

In addition to the statistical evidence, all of the questions are reviewed to make certain that they are appropriate for all of the people who will take the test. These reviews are carried out by specially trained reviewers at ETS as well as by committees of educators outside of ETS.

Does the SAT predict the freshman college grades of women as well as it predicts those of men? Actually, the correlations between college grades and SAT scores tend to be higher, on average, for women than for men. In that sense, the SAT is a slightly better predictor for women than it is for men.

One fact that makes the SAT appear to be biased against women is that women obtain average grades in college that are higher than the average grades of men, in spite of the women's lower average test scores. To resolve that paradox it is necessary to examine evidence about the criterion itself, the grades of women and men and the courses in which those grades are achieved.

Women tend to take more courses in college in which the average grades for the course are high. More men than women take courses such as calculus and physics in which fewer high grades are given. More high grades are given in courses in the humanities and social sciences which tend to enroll more women than men.[5]

Evaluating the validity and fairness of a test used for prediction requires an examination of the meaning of the variable that is being predicted. This is another example of the need to go beyond mere differences in scores to evaluate all of the evidence that bears on the validity and fairness of a test.

What Should We Do About Group Differences in Test Scores?

First, we should *not* automatically assume that all differences in average test performance are caused by bias in the tests. We live in a society in which there are still group-related differences in family income and opportunities for learning, both in and out of school. Young women and men still differ in interests, activities, and types and levels of courses taken. The qualities of schools that children attend are related to family income and place of residence. To blame the tests for the differences found in educational attainment is to ignore reality.

Second, we should *not* automatically assume that all tests are valid and fair for all of the inferences that are made on the basis of the scores. We should demand evidence that the test is meeting its intended purpose for all groups of examinees. The bulk of this paper has been devoted to explaining the various types of evidence that are required.

---

[5] Rigol, G. "Why Do Women Score Lower Than Men On The SAT?" *College Prep, Number 4.* New York: College Entrance Examination Board, 1989.

One might judge that if it takes so much evidence to be sure a test is valid and fair, why bother? Why not just quit using tests? Note, however, that the same evidence would be required to give us equal confidence in the fairness and validity of any other information on which we made corresponding inferences (e.g., grades, teacher judgments, letters of recommendations). We have subjected tests to a higher standard of evidence than many less formal measures. Not having the evidence on the other measures just allows us to ignore some of the difficulties and complexities of validation, not solve them. If we did not use tests, we should be asking all these same complicated and difficult questions about any measures we used in their places.

Third, we should use the information provided by fair and valid tests and other fair and valid measures to help improve education. Differences in indicators of educational achievement are a painful reminder that our goals of equality of opportunity have not been met. Such results can help us to pinpoint areas of greatest need and can help us to monitor our progress. They *should* be used to put pressure on the public to support better educational opportunities and better education as well as on the educational system to deliver a positive educational experience to all groups in this richly diverse nation.

References

American Educational Association, American Psychological Association, and National Council on Measurement in Education (1985) *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Berk, R.A. (ed.), *Handbook of methods for detecting test bias.* Baltimore: John Hopkins University Press, 1982, pp. 9–30.

Cole, N.S., and Moss, P.A. (1989) Bias in Test Use. In R.L. Linn (ed.) *Educational Measurement* (3rd edition) New York: American Council on Education/MacMillan. pp. 201–19.

Jensen, A. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–123.

Rigol, G. (1989) "Why Do Women Score Lower Than Men on the SAT?" *College Prep, Number 4.* New York: College Entrance Examination Board.

Wigdor, A., and Garner, W. (eds.). (1982) *Ability Testing: Consequences, and Controversies.* Washington, DC: National Academy Press.

Bias in Educational and Employment Testing: Selected Issues

By Lloyd Bond[*]

Introduction

In this paper I describe some of the major issues in educational and employment testing, review some of the procedures that have been advanced to detect and minimize possible biases in testing, and in the last section, respond to a series of specific questions posed to the panelists at the U.S. Consultation Meeting on the Validity of Testing in Education and Employment, June 19, 1989, Washington, D.C., sponsored by the U.S. Commission on Civil Rights. Space and time limitations do not allow a thoroughgoing discussion of the many issues involved. I have necessarily omitted discussion of many technical issues and glossed over others that could easily consume volumes by themselves. Where appropriate, references are given for more detailed discussions.

The Nature of Bias

In educational and employment contexts, a test may be biased in three major ways. First, a test is said to be biased if it purports to measure the same or similar attributes in different subpopulations of examinees (blacks vs. whites, males vs. females, etc.), but in fact measures different attributes depending upon the subpopulation. An example taken from Bond (1981) will illustrate this point. Suppose an eighth grade teacher wished to assess the verbal analogical reasoning ability of her class, which consisted of students from both urban and rural areas. Further, imagine that the test consisted largely of words that persons raised on a farm would be intimately familiar with, but that persons raised in the city would be less familiar with. A typical item on the test might be:

pig : sty :: chicken :

a) dinner b) plow c) turkey d) coop e) barn

It should be obvious that students raised on a farm would be at a tremendous advantage over others on items such as these. Rural students are much more likely to be familiar with the words comprising the analogy and hence are more likely to deduce their relationship to each other. Urban students will score lower on the test not because they are less proficient in analogical reasoning, but because they do not know the meaning of the words contained in the analogy. Under such circumstances, verbal analogical reasoning is confounded with vocabulary. The test is a *purer* measure of analogical reasoning for rural students because it is unconfounded with large differences in vocabulary. For urban students, one is never sure whether a low score represents unfamiliarity with the words comprising the analogy, or whether a low score indicates a more fundamental inability to reason analogically.

---

[*] Comments are welcome and may be addressed to Lloyd Bond, Department of Educational Research Methodology and Center of Educational Research and Evaluation, University of North Carolina at Greensboro, NC 17410.

Tests that measure one construct in one subpopulation and a different construct in another subpopulation are said to have *construct or categorical* bias. That is, the test, taken in its entirety, is biased against a given group or group(s) because it confounds the measurement of one construct with another. The only valid comparisons in the above example is the comparison of one rural student with another. All other comparisons are suspect. Comparing a rural student's performance with that of an urban student is obviously confounded with vocabulary differences. Comparing two urban students' performances, while less flawed, is nevertheless problematic because the differences in incidental knowledge of rural terms are likely to be greater in this group.

A second, related form of internal bias, known as *item bias* or *differential item functioning* (DIF) exists when only some of the items in a test work to the disadvantage of particular subpopulations of examinees, while other items are considered equally valid and appropriate for all groups. The statistical and methodological procedures used to detect such items have received considerable attention in recent years from measurement specialists.

The final way in which a test may be biased is in its ability to predict later performance on some activity of interest (e.g., performance in school or on the job). A test is said to be biased in this sense if in using scores on the test to predict later *criterion* performance, there result systematic errors of over or underprediction for one or more subgroups of examinees (Cleary, 1968). As with item bias, selection and prediction bias have been the subject of intense research and debate among measurement specialists over the past two decades. Procedures for investigating and/or attempts to minimize item bias and selection bias will be briefly reviewed. First, however, an important distinction, that between *bias and adverse impact,* concepts often confused in public debates, needs to be clarified.

Bias and Adverse Impact

Adverse impact exists wherever observed score differences between groups, *whether they reflect genuine, valid differences or not,* result in decisions that adversely affect one of the groups. Thus, tests that place minority youngsters in classes for the educably mentally retarded (where such placement is considered educationally harmful), tests that result in proportionally greater numbers of minority applicants being denied teaching certificates, and tests that reject minority job applicants disproportionately are said to have adverse impact. It is the mean score difference and its consequences that defined adverse impact. Bias, on the other hand, exist only when group score differences do *not* represent genuine differences in the construct being measured. An example may serve to clarify this important distinction. Suppose an employer has openings for a job that requires significant upper body strength. A test designed to measure the minimum upper body strength necessary to safely and effectively perform the job would probably result in the selection of disproportionately small numbers of women. In this situation, the test is not biased against women, but has significant adverse impact on women because it results in the selection of disproportionately small numbers of women. A test may have adverse impact on a group without necessarily being biased.

Analytical Procedures for Investigating Item Bias

Statistical approaches to the detection of potentially biased items are internal methods (Shepard, 1981) that assume the test *in general* is valid for all groups of examinees. These methods seek to find particular items that are troublesome. Statistical item bias techniques cannot aid in the determination of pervasive or categorical bias.

As Shepard (1981) has noted, the strength of such methods resides in the availability of multiple items all designed to measure the same thing and all analyzed separately. The methods are quite useful in helping to detect distortions or differential meaning in what was thought to be a homogeneous set of items.

In discussing the various internal procedures to detect bias items that have been proposed, I will attempt to keep the discussion as nontechnical as possible. In doing so, many methodological niceties will be omitted. A more complete technical discussion of these methods can be found in Berk (1982), and Wainer and Braun (1988).

Plotting Methods. This formerly popular approach, due to Angoff (1972), requires that item "p-values" or proportion correct for each item be calculated for each group of interest. By assuming that the attribute being measured is normally distributed in all relevant subpopulations of examinees, the p-value for each item is first transformed to a percentile scale and is then linearly transformed to have a mean of 13 and a standard deviation of 4. This scale transformation results in the "delta scale" used by the Educational Testing Service to indicate an item's difficulty level. The relative difficulty of the items for two groups of examinees may be compared by forming a bivariate plot of the delta values, with the x-axis representing item difficulties for one group and the y-axis representing the item difficulties for the other group. The resulting scatter plot of items normally forms an oval or ellipse, and items that have equal *relative* difficulty within each group will fall along the major axis of this ellipse. Aberrant items (that is, items suspected of being biased) are those which deviate from the major axis of the ellipse by some prespecified amount.

Items below the major axis are relatively more difficult for the x-axis group and items above the major axis are relatively more difficult for the y-axis group. As with all of DIF procedures, items that deviate from expectation by some prespecified amount are flagged and reviewed for possible clues to the source of the problem.

The major shortcoming of the Angoff approach to detecting biased items is that items that *genuinely* distinguish between high and low scorers on the test will be flagged as possibly biased by the procedure, when in fact true differences exist between the two groups being compared. It is for this reason that the Angoff procedure for detecting differential item functioning is no longer used by most researchers.

Chi-Square Methods. Chi-square methods (Scheuneman, 1979) assume that an item is unbiased if the probability of a correct response for individuals at comparable ability levels is the same regardless of their group membership. The name of the procedures stems from the use of the familiar chi-square test of goodness-of-fit to test whether persons from different subpopulations who have been "equated" on ability, have the same probability of getting a given item correct. Thus, an item is presumed to be unbiased if the proportion of individuals at any one ability level (regardless of group membership) who get the item right is the same.

In chi-square procedures, the test score range is normally divided into quintiles (the first quintile is defined by that point on the score scale below which 20 percent of the entire population of examinees fail, the second quintile is defined by the 20th percentile score and the 40th percentile score, and so on). It is assumed that persons in the same quintile are of roughly comparable ability. More exact control for ability can be attained by dividing the score scale into smaller and smaller percentile groups. If a given item is unbiased, the proportions of any two groups in a given quintile who get the item correct should be the same for both groups. (It is particularly important to keep in mind that the chi-square methods do not require that all subpopulations have the same proportion of examinees in each quintile. Rather, *of those persons in each group who are in a given quintile, the proportion getting a given item right should, within sampling error, be the same across all groups.)*

More recently, researchers at the Educational Testing Service (Holland & Thayer, 1986) have proposed using the Mantel-Haenszel procedure (Mantel & Haenszel, 1959) to detect items that function differentially across groups. This is also a chi-square procedure, but instead of dividing the score scale into gross percentile groups, the score scale is divided into every possible score. Thus for a 50 item test, there are 51 possible score groups (O to 50 inclusive). The procedure handles small cell frequencies by differential weighting, cells with larger frequencies receiving proportionately larger weights than cells with smaller frequencies. A detailed description of this promising approach to detecting biased items can be found in Wainer and Braun (1988).

Item Response Theory Methods. The most technically sophisticated and elegant approach to detecting biased items in a test are based upon a model of testing known as item response theory (IRT), developed in the early 1950s by the eminent psychometrician, Frederick M. Lord. Because of the complex estimation procedures involved, IRT did not become a popular model for test development until the advent of high-speed computers in the 1960s. IRT is now the preferred test model for numerous testing applications including computer adaptive testing, test equating, item banking, and item bias research. A detailed discussion of the IRT approach to the detection of biased test items can be found in Lord (1980). Only a sketch of the procedure is given here.

Two basic assumptions underlie item response theory. The first is that ideally, all items on a test measure one and the same attribute and no others. The second assumption is that each item is a *separate and independent* measure of the attribute, so that knowing the answer to any one item does not aid the examinee in answering any other item on the test.

Another important feature of IRT involves the estimation of a person's ability. In ordinary test scoring, a person's total number of right answers (sometimes corrected for guessing) is taken as the estimate of his or her ability. This is not so in IRT. Rather, examinee ability is estimated via a complex, iterative procedure that depends upon a person's pattern of right and wrong answers. It should be noted that, if the assumptions underlying this test model are met even approximately, IRT represents a powerful advance over traditional approaches to ability measurement. The reason for this is that, unlike regular test scoring, estimates of an individual's ability using IRT does not depend upon *the specific* items included in the test, nor does it depend upon the group to which the person is being compared. In this sense, "scores" derived from the IRT model are more nearly absolute than are scores derived from traditional measurement methods.

With IRT, each item may be represented by a graph or curve which shows the various probabilities of getting the item right as a function of the examinee's increasing ability. An item is flagged as possibly biased if the two curves for the majority and minority group differ by more than what would be expected from mere sampling error. In this sense, IRT is similar to the chi-square methods: persons of equal ability should have the same probability of getting any given item correct.

Models of Selection Bias

Selection bias has been the subject of intense methodological debate and policy interest. While complete unanimity in the psychometric community as to the most effective way to handle selection bias and adverse impact has not been achieved, much has been learned about the pluses and minuses of various approaches. The various selection models along with their advantages and disadvantages will be briefly reviewed. Detailed analyses of their strength and weakness may be found in Jensen (1980), and the 1976 special issue of the *Journal of Educational Measurement,* (volume 13, no. 1).

Some 15 years ago, Petersen and Novick (1976) provided one of the most comprehensive and penetrating discussions of the various models of selection and prediction bias. Their terminology and analysis have generally become the standard of the profession. Because of the central importance of the regression model in understanding other models of fair selection, this method will be discussed at some length.

The Classical Regression Model. According to Cleary (1968):

> A test is biased for members of a subgroup of the population if, in the prediction of the criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low (p. 115).

This definition of test bias is called the regression model because in practical situations the existence of bias is determined by examining the least-squares linear regression lines (where the vertical axis is the criterion performance and the horizontal axis is the test score) for two different groups. If the regression lines differ in (1) their slopes or (2) where they intercept the y-axis, then the test is biased according to this definition.

Figure 1 illustrates, for two hypothetical groups labeled 1 and 2, three situations where bias would be said to exist according to the classical regression model, and a fourth situation where no bias exists. In the figure it is assumed that the range or spread of scores on the test and the range of criterion performance is roughly equal for both groups, although the means will generally differ. Figures 1a and 1b are examples of *intercept bias.* The term "intercept bias" stems from the fact that the relationship between the test and the criterion (called the slope of the regression line) is the same for both groups, but the regression lines differ in where they intersect the y-axis (performance).

In figure 1a, the two groups are indistinguishable on the criterion, but differ significantly on the test. Thus, two individuals, one in group 1 and one in group 2, who perform identically on the criterion, have significantly different test scores. The test is biased against members of group 1. Members in the shaded portion of group 1, who would have been successful, are rejected in favor of members in the shaded portion of group 2, who were in fact unsuccessful.

In figure 1b the opposite occurs. The groups are indistinguishable on the test, but differ significantly on their criterion performance. The slopes of the regression lines are identical; but, for a given test score, the criterion performance for members of group 1 is systematically higher than that for members of group 2. Under this model, then, the test is biased against members of group 1 since many in this group who would have been successful are rejected in favor of many members of group 2 who in fact proved unsuccessful.

Figure 1c illustrates *slope bias or differential predictive validity* according to the regression model. Here, the strength of the relationship between test scores and criterion performance differs for the two groups. For group 2, the relationship between test scores and job or school performance is strong. Two individuals in this group who have widely different test scores also have widely different levels of performance on the criterion. By contrast, in group 1, widely different scores on the test correspond to only modest differences in criterion performance. The test is a valid predictor of job or school performance for group 2, but is far less valid for group 1.

In Figure 1d there is no bias according to the regression model. Note that the regression line for group 1 and the regression line for group 2 are one and the same. There is neither consistent underprediction nor overprediction for either group and the strength of the relationship between test and criterion (that is, the slope) is the same for both groups. Members of group 1 are low on the test, but also do less well on the criterion. Members in group 2 score high on the test, but have correspondingly high performance on the criterion.

The regression model described above has the reputation among the majority of measurement specialists as a psychometrically sound model of fair selection. Its straightforward application, however, generally results in few minority applicants being hired compared to their percentage of the applicant pool. This circumstance has sparked a number of alternative models.

The Proportional Representation Model. This model specifies that the proportion of applicants selected from the majority and minority group should reflect their respective percentages in the applicant pool. Although this model has much support among those who believe tests are categorically biased against minorities, it has much less support among measurement specialists because it assumes beforehand that score differences are the result of bias in the test.

The Equal Risk Model. The Equal Risk Model, first described by Einhorn and Bass (1971) specifies that a test is fair if it selects applicants, regardless of group membership, in order according to their risk of failing below the minimum acceptable performance. Jensen (1980) correctly points out that if the regression lines for minority and majority applicants are equal

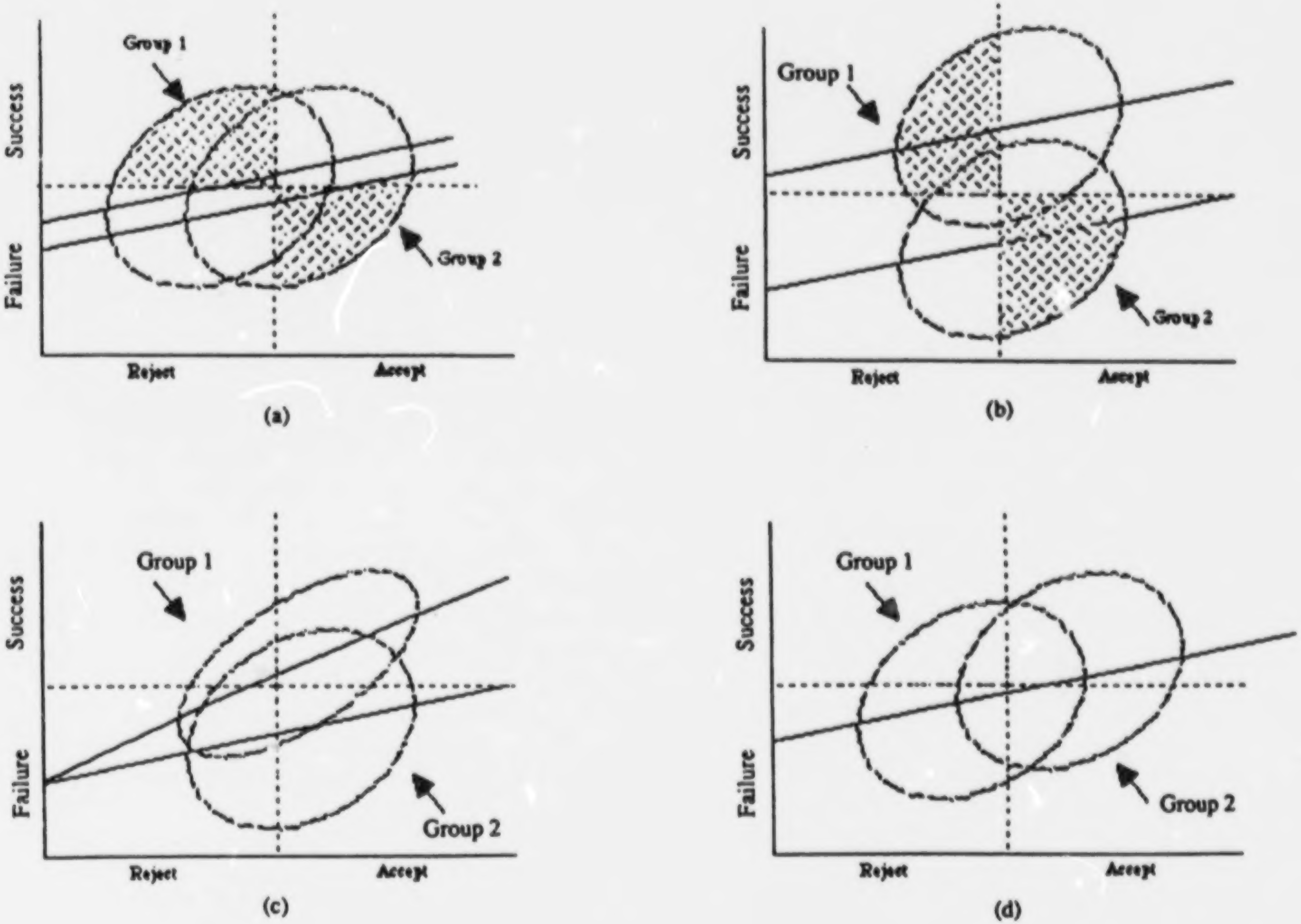
Group 1
Success
Failure
Group 2
Reject
Accept
(a)
Group 1
Success
Failure
Group 2
Reject
Accept
(b)
Group 1
Success
Failure
Group 2
Reject
Accept
(c)
Group 1
Success
Failure
Group 2
Reject
Accept
(d)

Figure 1: Examples of Selection Bias (a,b,c,) According to the Regression Model

and if the precision with which the test predicts performance for the two groups in the same, then this model is identical in its results to the regression model. If, however, the precision with which the test predicts the performance of minority and majority groups differ (that is, if the standard error of estimate differs for the two groups), then the Equal Risk Model diverges from the Regression Model and may select persons with lower predicted performance over those with higher predicted performance.

The Constant Ratio Model. The Constant Ratio Model was first proposed by the R. L. Thorndike (1971) and specifies that cut score(s) on the selection test should be set such that applicants from any two groups are selected in proportion to the fraction of the two groups reaching a specified level of criterion performance. The rationale underlying this model stems from the fact that, in practice, with imperfect tests (as all tests are) it often nappens that the difference between the majority and minority group means on the test is greater than their difference on the criterion. When this happens, proportionally more minority applicants who would have been successful are rejected compared to majority applicants. This fact is also the basis for the recent recommendation by the National Academy of Science Committee for the continued use of within race norming of the General Aptitude Test Battery (GATB).

The Conditional Probability Model. The Conditional Probability Model, Cole (1973), states that for both minority and majority groups whose members can achieve satisfactory criterion performance, there should be the same probability of acceptance regardless of group membership.

The Equal Probability Model. The Equal Probability Model, first describe by Linn (1973) as an alterative logical possibility, rather than as a model to which he subscribed, specifies that the cut scores for the majority and minority groups should be set so that the proportion of selected persons predicted to succeed on the criterion is the same for both groups.

Unlike the Regression and Equal Risk Models, the Thorndike, Cole, and "Linn" models can and most often will result in different cut scores being set for the majority and minority groups. The appeal of the models lies in the fact that in most situations likely to be encountered in practice, they will result in more minority applicants being hired than would be the case under the Regression or Equal Risk Models. The models thus advance a socially desirable goal. The three selection strategies have been criticized, however, by Petersen and Novick (1976) on the grounds that they may discriminate against certain minorities (e.g., Japanese Americans) and on the grounds that they are "internally inconsistent." That is, the models are concerned with fairness to those who pass the test or who would be successful on the job. If one extends the notion of fairness to include those who failed the test or those who would not succeed on the criterion, then different cut scores have to be set. A single cut score cannot satisfy both conceptions of fairness. Thus, a "Converse" Constant Ratio Model assumes a selection procedure is fair if cut scores are set so that the proportion rejected compared to the proportion unsuccessful is the same in the minority and majority group.

Models based on Expected Utility. Petersen and Novick (1976), Gross and Su (1975) and others have advocated a model based upon classical "utility" theory. This approach to fair selection maintains that the affected parties (employers, minority groups, the public generally) must eventually come to consensus on the *value* they attach to certain outcomes. For example, the distaste for false positives (accepting someone who will fail) must be weighed against the desirability of increasing the pool of minority doctors, police officers, and teachers. The desirability of making correct decisions (accepting applicants who turn out to be successful or rejecting applicants who would have failed) must be weighed against the undesirability of decreasing even further the numbers of minorities in certain occupations. If (and it is a big "if") consensus can be achieved, utility models attempt to quantify this consensus judgment and to set cut scores so as to maximize the overall desirability of the outcomes of the selection process.

Issues

1. How should biased items be identified?

The converging evidence, from analyses of real tests as well as from analyses of tests simulated to include biased items, suggests that the Mantel-Haenszel procedure and the IRT-based approaches are superior to other approaches in identifying items that behave differently across subpopulations of test takers.

2. Should biased items be categorically eliminated? What factors govern this choice?

There are two schools of thought on this issue. The first is that no strictly statistical procedure should govern, exclusively, the inclusion or exclusion of an item from a test. According to this view, statistical procedures for identifying biased items are useful only as *aids* to professional judgment. They serve merely to alert test developers to possible flaws in wording, distractors, and so on, that may have been overlooked earlier in test development. Hence, according to this view, it is entirely possible that an item flagged by a statistical detection procedure is ultimately determined to be psychometrically sound. The professional consensus may be that the relative differential difficulty of the item cannot be traced to irrelevant characteristics of the item. If the item is judged valid on content and predictive grounds, then this school of thought maintains that the item should be retained even though it may be "biased" in the statistical sense.

The second school of thought (and one to which I now subscribe) maintains that the same decision rules about item inclusion that are used in traditional item analysis should be used here. In traditional test development, for example, an important statistic that is used to judge the quality of an item is the item's correlation with all of the other items on the test. If those examinees who get the item right do no better on the total test than those who get the item wrong, then the item is discarded from the final form of the test because it only increases the errors of measurement in the test. Even if the professional judgment is that the item is otherwise sound, it is sill discarded. (It should be noted that measurement specialists who subscribe to the first school of thought agree with this procedure.) I believe the level of technical development in item bias research is now such that these procedures should be included routinely in test development along with other traditional item analysis statistics. As such, items identified by the best procedures as biased should be altered and, if this is not successful, the item in question should be removed from the final form of the test.

3. What proportion of items in current tests are biased?

This is a difficult question that is probably not answerable at this time. I am most familiar with DIF analyses of the SAT, and even here I am not certain how many items are judged to be biased. A conservative estimate is that between 5 and 10 percent of the tryout items for established testing programs are flagged as potentially biased by the better statistical procedures.

4. How much does eliminating items with DIF reduce group differences?

The answer to this question depends, among other things, upon the number of items eliminated, the overall difference in the proportion of the subpopulations of interest who correctly answer the item, and the item's correlation with the total test score. It also turns out that sometimes items are identified that are biased against the *majority* group. If these items are also eliminated, then the overall effect on group differences will of course be lessened. Again, the only popular testing program with which I am familiar that is routinely using DIF procedures in item analysis is the SAT. The results so far indicate that the reduction of group differences tends to be small.

5. Is there differential predictive validity for black/while, male/female, etc.?

As a general rule, correlations between test scores and school performance, and correlations between test scores and on-the-job performance are not substantially different for males and females and blacks and whites.

6. How high should correlations be for a test to be valid?

It is probably not advisable to ask this question in terms of "correlations," since correlations can be low and still the test can be useful, and vice versa. It is better to frame the question in raw regression terms, that is, in terms that allow one to say that an increase in test scores from X to X' corresponds to a predicted increase of Y to Y' in nondefective pieces produced, dollars of sales volume, and so on. When stated in this fashion, the question is best answered by the affected parties, not by a measurement specialist.

7. If predictive validity is high across groups, is it necessary to obtain other forms of validity (content, job relatedness, etc.) as well?

Categorically yes. Predictive validity may be high for a whole host of wrong reasons. To take a deliberately extreme example, consider a job for which it has been demonstrated that a battery of cognitive tests are good predictors. Since whites as a group tend to score higher on standardized tests than Hispanics and blacks, an employment test based upon skin color alone could have a modest, possibly significant correlation with job performance! Of course, such a blatantly racist selection procedure should never be used, but it does point out the fundamental flaw in relying solely on predictive "validity" evidence. Other forms of validity are absolutely essential. To take but one of many real-world examples, virtually all paper-and-pencil tests contain an inflated verbal component that may not be related to job performance, but is related through educational differences to the cognitive abilities that *are* job related. Without some statistical adjustment for the vitiating effects of verbal ability, persons with little formal education who are good mechanically, for example, would be penalized if only predictive validity were used. Moreover, where supervisory ratings are the criterion, it is conceivable that persons who are good verbally might receive high performance rating for reasons totally unrelated to job performance, *per se*.

To see more clearly why reliance on predictive validity alone is not sufficient, consider the following hypothetical, but possibly common, situation. In "double blind" predictive validation studies, applicants are hired without consideration of their test scores, and neither the personnel researcher nor the supervisor knows the test scores of the validation sample of workers. Job performance is then later correlated with test scores. If supervisory ratings are used as part of the measure of criterion performance, then *prejudicially* low ratings of black employees who as a group probably scored lower on the test would result in an inflated predictive validity coefficient for the test. Hence, it is possible to obtain an *erroneously* high predictive validity coefficient even when the validation procedure satisfies the research "ideal" (i.e., the double-blind procedure).

8. How should job analysis and content validation be done? Who should do this?

(In attempting to answer this question, I should state first that it is surprising how the criterion, being such an integral part of the evaluation of the predictive validity of a test, has historically been one of the weakest links in the validation chain.) A relatively detailed description of job analysis methods can be found in Landy (1985). To paraphrase Landy (1985), there are really only three ways to get information about the elements that make up a job: ask someone about the job who knows it well, watch a competent incumbent carry out the tasks that comprise the job, or try to do the job oneself. The latter is rare and impractical and will not be discussed further. Far and away the most common method for conducting a job analysis is a combination of interviews and questionnaires. Typically, the job analyst reads as much about the job as possible and then interviews competent incumbents and supervisors. A list of the most important and frequently occurring job tasks is then developed with the aid of incumbents and supervisors. The list is then reviewed by many other incumbents as well. The two dimensions of *importance* and *frequency* are the essence of a competently conducted list of job elements.

Because of their efficiency and low cost, interviews and questionnaires are the method of choice in most job analyses, but, alone, they have their weaknesses. First, there is the simple reality that many incumbents and supervisors may be suspicious, busy, or both. In addition, expert performers are often unaware of how they carry out their duties, or are unable to describe them accurately. This is especially so for tasks that have become so routinized and habitual that they are "second nature." It is for these reasons that a good job analysis should include actual observation of competent workers performing the job. To be sure, this is an expensive proposition, for it can involve actually going on the beat with a police officer or accompanying a plant supervisor throughout a typical day. But the information gained from expert observation can be invaluable in accurately specifying the content of a job.

9. & 10. What are the legal and policy issues relating to the development and use of tests? What effects have legislation, litigation, and government regulations had on testing?

The legal and policy issues relating to the development and use of tests in education and employment and the effects of legislation, court action, and government regulations on testing have been very competently summarized by Rudert (1989) in the her Background Paper to the Consultation Meeting of June 16, 1989, in Washington, D.C. By way of update, I would only add that since her discussion, the courts have moved even further away from employers' obligation to justify discriminatory impact, and further in the direction of plaintiff's obligation to prove discriminatory intent. Advocates for minority causes have complained, justifiably in my opinion, that individual citizens simply do not have the financial and administrative wherewithal to successfully gain legitimate relief under such circumstances.

11. What influence has social science had on legal and regulatory processes?

I quite agree with Rudert's (1989) observation that "the flow of information between social scientists and those who make law (and vice versa) is uneven" (p. 35). On the one hand, while the courts have relied on "expert testimony" in litigation involving placement in special education classes and minimum competency testing, more often than not, such expert testimony has tended to reflect the philosophical opinions of the witnesses, rather than a hard and fast fidelity to research-supported data. The abysmal state of affairs is perhaps nowhere more clearly seen then in the approach taken in *PASE*, where the presiding judge, frustrated by the conflicting expert opinion, took it upon himself to decide via purely subjective examination which items on the Stanford-Binet were biased against black children and which were not.

The one encouraging connection between the social sciences and legal processes is the increasing reliance of the courts on the APA/AERA/NCME *Standards for Educational and Psychological Testing*, especially in teacher certification and employment testing.

12. What legal or regulatory changes should be made with respect to testing?

Many suggestions for how Federal or State agencies should regulate testing have come in and out of favor. Some have advocated a national truth in testing law that would require complete disclosure of all item development procedures and supporting data for any publicly mandated, nonvoluntary test, and for any test used for professional certification or admission to higher education. George Madaus of Boston College has been a particularly eloquent spokesman for an advisory committee composed of measurement specialists, public officials, and relevant affected parties to monitor the use of testing in American society. I believe this is a move in the right direction that could have enormously beneficial consequences.

References

Angoff, W.H. (1972). A technique for the investigation of cultural difference. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Berk, R.A. (1982). *Handbook of methods for detecting test bias.* Baltimore: The Johns Hopkins University Press.

Bond, L. (1981). Bias in mental tests. In B.F. Green (ed.), *New directions in testing and measurement: Issues in testing, coaching, disclosure, and ethnic bias.* San Francisco: Jossey-Bass.

Wainer, H., & Braun, H. (1988). *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–24.

Gross, A.L., & Su, W. (1975). Defining a "fair" or "unbiased" selection model: A question of utilities. *Journal of Applied Psychology, 60,* 345–51.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (eds.), *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Jensen, A.R. (1980). *Bias in Mental Testing.* New York: Free Press

Landy, F.J. (1985). *Psychology of Work Behavior.* Homewood, IL: The Dorsey Press.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Mantel, N., & Haenszel, W. (195). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–48.

Rudert, E.E. (1989). The validity of testing in education and employment. Background Paper for Consultation. Washington, DC: U.S. Commission on Civil Rights.

Scheuneman, J. (1979). A method for assessing bias in test items. *Journal of Educational Measurement, 16,* 143–52.

Standardized Testing: Harmful to Civil Rights

By D. Monty Neill
National Center for Fair and Open Testing (FairTest)[*]

In the last two decades, standardized multiple-choice tests have come to dominate the educational landscape in America. From pre-school to college, these exams have become major criteria for a wide range of school decisions. Test scores limit the programs students enter and dictate where they are placed; standardized exams determine the shape of the curriculum and the style of teaching; and their results are used to assess the quality of teachers, administrators, schools, and whole school systems. Across the Nation, standardized, multiple-choice exams are increasingly required before candidates can be certified as teachers. Their use for licensure, or as a means of sorting job applicants, is also widespread in other occupations.

Taken as a whole, tests are one of the Nation's most important gatekeepers for social mobility and advancement from pre-school through employment. But, rather than enhancing equity and enabling access, tests have become unfair barriers that have destructive effects on equal opportunity, educational quality, and the Nation's economy.

Hundreds of Millions of Tests

A recent study by the National Center for Fair & Open Testing (FairTest) estimated that public schools in the United States administered 105 million standardized tests to 39.8 million students during the 1986–87 school year. That is an average of more than two and one-half tests per student per year. At that rate, by the time a student graduates, he or she will have taken 30 standardized tests. Virtually all are multiple-choice and machine-scorable.

The annual total includes over 55 million standardized achievement, competency, and basic skills tests administered to fulfill local and State testing mandates. An additional 30 to 40 million tests were given to compensatory and special education students. Two million more tests were used to screen kindergarten and pre-kindergarten students, and 6 to 7 million college and secondary school admissions, General Equivalency Degree (GED) and National Assessment of Educational Progress (NAEP) tests were administered that year.[1]

This estimate of 105 million tests per year is conservative. The total does not include tests administered to identify or place "gifted-and-talented" or limited-English proficient students, for which there are no reliable figures. Nor does it include tests administered by private and parochial schools to their students. Moreover, the FairTest survey counted each administration of a test battery as only one test, but some included up to five separate exams. Thus, the total could be double the initial estimate.

---

[*] Thanks to Noe Medina of Education Policy Research and the staff of FairTest for substantial help on this paper. © FairTest 1989.

[1] N. Medina and D.M. Neill. *Fallout From the Testing Explosion.* (Cambridge, Mass: FairTest, 1988). Some of the material in this article is elaborated in the Medina and Neill report, which also contains an annotated bibliography.

The FairTest survey also revealed that the number of States that mandate school testing has increased greatly in recent years. In addition, FairTest found that testing is most prevalent in the southern States and in large urban school systems. Both tend to have higher percentages of low-income and minority students than the national average.

The survey did not count standardized tests administered to college and university students after enrollment, an area which is rapidly growing. Many of these tests act as barriers between 2- and 4-year institutions or lower and upper level programs. Nor did the survey tally exams administered for licensure or employment by government agencies (e.g., civil service tests) and private employers. While uncounted, these likely number in the tens of millions annually.

Test proponents, of course, applaud these trends. They see tests as "valid" and "objective" mechanisms to inject "accountability" and thereby improve student achievement, educational quality, and employee competence. Not surprisingly, standardized exams have been an essential element of the "School Reform Movement."

Experience with standardized test use in education, however, paints quite a different picture. Rather than being "fair" and "objective" instruments, standardized tests often produce results that are inaccurate, inconsistent, and harmful to minority, low-income, and female students. By narrowing the curriculum, frustrating teachers, and driving students out of school, overreliance on testing undermines school improvement instead of advancing its cause. Rather than promoting accountability, the testing frenzy shifts control and authority into the hands of an unregulated testing industry. As a result, using standardized test scores as the primary criterion for making important educational decisions has led to less public understanding of the schools and a weaker educational system.

Standardized employment tests are no more "objective" than educational tests. Ample evidence demonstrates that they exclude many qualified applicants, a disproportionate number of whom are minorities. In addition to causing often irreparable harm to the applicants who fail, they also hurt the industries in which they are used by excluding potentially valuable employees.

As the population of the U.S. diversifies, the economic well-being of the Nation requires that minorities no longer be excluded by arbitrary barriers. The social health of the Nation likewise is endangered when education and employment opportunities are undermined by testing, consigning minorities to continued disproportionate placement at the lowest socioeconomic levels.

Alternatives to the misuse and overuse of standardized tests do exist. Appropriate, authentic assessment methodologies have been developed that avoid many of the problems of standardized exams. These alternatives should be disseminated and implemented to largely replace standardized tests with fairer and more helpful assessments.

The Inadequate Quality of Standardized Tests

Standardized tests are consistently sold as scientifically developed instruments which simply, objectively, and reliably measure achievement, abilities, or skills.[2] In reality, the basic psychological assumptions undergirding the construction and use of standardized tests are open to question. Studies conducted to determine test reliability and validity are often inadequate. Many tests are administered in environments that contradict claims of "standardization."

These flaws undermine test makers' claims of objectivity and often produce test results that are inaccurate, unreliable and ultimately invalid. As a result, tests generally fail to effectively and usefully measure test takers' achievement, abilities, or skills.

False Assumptions

The ability of standardized tests to accurately report knowledge, abilities, or skills is limited by assumptions that these attributes can be isolated, sorted to fit on a linear scale, and reported in the form of a single score. Gould labels these the fallacies of *reification* (e.g., treating "intelligence" as though it were a separable unitary thing underlying the complexity of human mental activity) and *ranking* ("our propensity for ordering complex variation as a gradual ascending scale"). He concludes, "(T)he common style embodying both fallacies of thought has been quantification, or the measurement of intelligence as a single number for each person."[3] This "style" also pervades achievement and ability testing.

Many of the assumptions and structures of achievement tests are based on IQ tests and operate in the same way. For example, assumptions regarding the unidimensionality of ability and development are common to both.[4] Such assumptions are at odds with contemporary research, which emphasizes diversity in the nature and the pace of child development.[5] In general, modern theories emphasize the complexity of human intelligence and ability. Researchers have observed that knowledge, learning, and thinking have multiple facets, and that a high level of development in one area does not necessarily indicate a high level of development in others.[6]

Test constructors not only erroneously presume that the knowledge, skill, or ability being measured is one-dimensional, but also that it tends to be distributed according to the "normal" bell-shaped curve. The bell-shaped curve is used for statistical convenience, not because any form of knowledge or ability has been proven to be distributed in this manner.[7] The use of a linear scale curve can result in tests labeling performance (ability or achievement) as incorrect or substandard when it is simply a normal variation; and it can mask real differences in ability or achievement by lumping attributes together.[8]

---

[2] As Levidow observes, deciding what to measure and what not to measure is a socially determined act. L. Levidow, "'Ability' Labeling as Racism," in D. Gill and L. Levidow, eds., *Anti-Racist Science Teaching* (London: Free Association Books, 1987).

[3] S.J. Gould. *The Mismeasure of Man* (New York: Norton, 1981.), 24.

[4] B. Singh. "Graded Assessments," in Gill and Levidow, eds., (1987).

[5] "NAEYC Position Statement on Developmentally Appropriate Practice in the Primary Grades, Serving 5-Through 8-Year-Olds." *Young Children* (January 1988).

[6] H. Gardner. *Frames of Mind: The Theory of Multiple Intelligences* (New York: Basic Books, 1985).

[7] C. Ryan. *The Testing Maze.* (National PTA: Chicago, Ill., 1979), p. 8.

Unitary test scores and linear scaling of scores ignore true human complexity and thus provide a deceptive picture of individual achievement, ability, or skills. This is a fundamental problem underlying standardized tests in education and employment.

Test Reliability

Claims that standardized tests exhibit a high level of reliability are usually taken to mean that test results will be similar in successive administrations. In fact, test "reliability" is a technical term which encompasses several different concepts.

The type of reliability generally measured and reported for standardized tests is internal or interform reliability. Consistency over time, which many would consider of greater importance, is infrequently measured and reported by test publishers. This type of study generally produces lower reliability coefficients and is more expensive to conduct.[9]

The level of test reliability (regardless of the type of reliability measured) is reported as a "reliability coefficient" on a scale from 0 to 1. For most standardized tests, the reported coefficients are high—often exceeding .8 or .9.[10]

Yet, for an "IQ" test with a reliability coefficient of .89 and a standard deviation of 15, a student has a reasonable likelihood of having a "true score" of up to 13 points higher or lower.[11] Thus a school system could, for example, deny entry into a "gifted and talented" program requiring an IQ of 130 to a student scoring 117 when that student's "true score" could well be 130.

Admission to college or employment may be denied for similar reasons. Many universities, particularly State institutions, have established cut-off scores on admissions tests. However, on the SAT, for example, due to the standard error of difference, two test takers' scores must differ by at least 138 points before the test maker is sure that their measured abilities differ. Nonetheless, even 10 points, just one question, may cause an applicant to be denied entrance, regardless of any other qualifications or evidence of capability.[12]

Due to nonstandard administration and examiner impact on the test taker, test administration procedures reduce reliability below the figures reported from experimental settings. Administration effects particularly harm low-income and minority school students. For example, black students are less apt to perform well with an administrator they do not know, while an anonymous administrator does not affect middle-class white children.[13]

---

[8] Medina and Neill (1988), p. 10. See also O.L. Taylor & D.L. Lee, "Standardized Tests and African-American Children: Communication and Language Issues," *Negro Educational Review* (April–July 1987), 67–80.

[9] *Ninth Mental Measurement Yearbook* (1985). See reviews of the California Achievement Test, Comprehensive Tests of Basic Skills, Iowa Tests of Basic Skills, Metropolitan Achievement Test, Stanford Achievement Test, SRA Achievement Series, and Gesell Preschool Test.

[10] A. Anastasi. *Psychological Testing* (sixth edition) (New York: Macmillan Publishing Company, 1988). See also reviews cited in note 9.

[11] Anastasi (1988). See discussion in chap. 5, esp. on "Standard Error of Measurement."

[12] *1988–89 ATP Guide for High Schools and Colleges* (Princeton: The College Board, 1988). H. Breland, G. Wilder, and N. Robertson. *Demographics, Standards and Equity: Challenges in College Admissions* (American Association of Collegiate Registrars and Admissions Officers, *et al.*, 1986).

Because reliability is often much lower for subsections of achievement tests and for tests administered to young children (below 9 years of age), the chance for error increases when decisions are made based on subtest scores or when tests are used for placing young children.[14] Cautions against such potential test misuses are often buried deep inside hard-to-read manuals.

In general, *no test has sufficient reliability to warrant making decisions solely or primarily on the basis of test scores.* Such decisions have been shown to disproportionately harm low-income, minority, and younger students. However, school systems, universities, and employers routinely make decisions on this flawed basis.[15]

Test Validity

"A test," write Airasian and Madaus, "is a sample of behaviors from a domain about which a user wishes to make inferences. . . . Test validity involves an evaluation of the correctness of the inferences about the larger domain of interest."[16]

Validity in standardized tests tells us whether a test measures what it claims to measure, how well it measures it, and what can be inferred from that measurement. Test validity cannot be measured in the abstract but can only be determined in the context of the specific uses to which a test's results will be put. Thus, information and conclusions regarding a test's validity in one context may not be relevant and applicable in different contexts. It is rarely an all-or-nothing proposition; rather, it is a process of accumulating evidence to justify use of a test in a given situation.[17]

Like reliability, the term "validity" encompasses several concepts:

---

[13] On the importance of standardized administration, see Anastasi (1988), 34 and 38. On lack of standardization in administration, see K. Wodtke, *et al.*, "Social Context Effects in Early School Testing: An Observational Study of the Testing Process" (paper presented at the 1985 American Educational Research Association Annual Conference), 28. For bias due to administration, see D. Fuchs & L.S. Fuchs, "Test Procedure Bias: A Meta-Analysis of Examiner Familiarity Effects," *Review of Educational Research* (Summer 1986), 243–62; "Test Conditions Can Harm Minority-Group Children," *The Chronicle of Higher Education* (Nov. 18, 1987), A15.

[14] *Ninth Mental Measurement Yearbook* (1985), see reviews cited in note number 10. See also, L.A. Shepard & M.L. Smith, "Flunking Kindergarten: Escalating Curriculum Leaves Many Behind," *American Educator* (Summer 1988), 36.

[15] In addresses to the "National Conference on the Technical Characteristics of National Norm-Referenced Achievement Tests" (The School Board of Palm Beach County, Fla., 1989), technicians from five testing companies repeatedly urged that tests not be used as sole criteria for decisionmaking.

[16] P.W. Airasian and G.F. Madaus. "Linking Testing and Instruction: Policy Issues," *Journal of Educational Measurement* (Summer 1983), 104.

[17] Anastasi (1988), ch. 6. American Educational Research Association, *et al., Standards for Educational and Psychological Testing* (Washington, D.C.: American Psychological Association, 1985), Part I.1.

— Content-related validity determines whether the test questions relate to the trait or traits or content domain the test purports to measure.

— Criterion-related validity compares test performance (for example, on a reading test) against a standard that independently measures the trait (such as reading ability) the test purports to measure. Criterion validity takes two forms, concurrent and predictive.

— Construct-related validity examines how well a test actually correlates with the underlying theoretical characteristics of the trait it purports to measure. For example, does the test accurately measure "academic ability" or "competence" or "reading"? This form of validity is rarely reported by test makers even though expert opinion has increasingly concluded that construct validity is the essence of validity.[18]

Content Validity. Content validity determines whether the test questions relate to the trait or content domain the test purports to measure. Multiplication questions on a test, for example, relate to the trait "ability to do multiplication" and thus would purport to measure knowledge of the content area of multiplication.

Consider, then, a test in U.S. history. The accumulation of items on the test is supposed to be an adequate proxy for the knowledge domain of U.S. history. A test taker who correctly answers a certain number of items will be said to have a corresponding level of knowledge about U.S. history.

The first question that arises is, "What are the items that should be on the test?" Items must be selected so that the test adequately covers the content domain. If the content domain is recall (names, dates, etc.), the test content can be correspondingly simple.

However, the domain is rarely so cut-and-dry as simple facts: no historian would reduce history to names and dates (however much it may so appear to many an unlucky student). Rather, history involves questions of methodology, relations among events, causes and effects, drawing conclusions from evidence, testing hypotheses, constructing theories, etc. And every one of these, from "facts" to theories, is subject to debate among historians.

A good history course, even prior to high school, will explore, at an appropriate level, the complexity that constitutes history. To be content valid at the level of sophistication of the appropriate domain, the test must cover what the domain covers. This is so difficult to do within the multiple-choice format that it is, essentially, not done. This format appears to be fundamentally incapable of measuring what are now commonly referred to as "higher order thinking skills."[19] Lack of adequate content validity can have wide-ranging effects. For example, if a U.S. history test only measures factual recall and the test is used to guide curriculum (as is increasingly the case), then not only will most of the real content of history not be measured, it will be excised from the curriculum.

---

[18] G. Madaus & D. Pullin. "Questions to Ask When Evaluating a High-Stakes Testing Program," *NCAS Backgrounder* (June 1987). Messick, S. "Meaning and Values in Test Validation," *Educational Researcher* (March 1989), 5–11. Messick, S. "The Once and Future Issues of Validity," in H. Wainer and H. Braun, *Test Validity* (Hillsdale, N.J.: Lawrence Erlbaum, 1988), 33–45. Cronbach, L., "Five Perspectives on the Validity Argument," in Wainer & Braun (1988), 3–18. Anastasi (1988), Chapter 6. *Standards* . . . (1985)

[19] N. Frederickson. "The Real Test Bias," *American Psychologist* (March 1984), 193–202. R. Marzano and A. Costa, "Question: Do Standardized Tests Measure General Cognitive Skills? Answer: No," *Educational Leadership* (May 1988), 66–71.

The selection of test items typically is done by panels of experts who review textbooks for content, draft, and then review items (a method occasionally referred to as BOGSAT: Bunch Of Guys Sitting Around a Table). Essentially, the subjective views of individuals are aggregated to design a test whose content is labelled "objective" and comprehensive.

The committee of experts must choose a set of questions that adequately represents the content domain. Each item must be one that reasonably should be on the exam, so experts are asked whether the item should be included. This is a simple, affirmative format.

However, what content validity studies need is the disconfirming hypothesis: What is not included? Is the overall balance of the items adequate to cover the content? Given the limited number of questions, is the content range a fair approximation of the domain?

Consider the case of the National Teachers Exam (the NTE), produced by Educational Testing Service (ETS). In many States, a prospective teacher must pass this test in order to obtain certification. The Core Battery of the test has three sections: General Knowledge, Communication Skills, and Professional Knowledge.

The assumption underlying the exam is that those who do not pass would not be good teachers. This predictive claim will be examined below. But it is also a content claim. For example, the Professional Knowledge test purportedly covers a representative and appropriate sample of the broad domain of basic professional knowledge.

However, in a study by the Rand Corporation, expert opinion was that "less than 10 percent of over one hundred questions required knowledge of theory, research or fact pertaining to teaching and learning." Questions about testing, however, were prominent, as were items about school law and administrative procedures, and items requiring agreement with the test makers' teaching philosophy, though their's is not the only philosophy of teaching.[20] It appears that the NTE Professional Knowledge test lacks basic content validity, perhaps because the chosen items apparently were never subject to disconfirming hypotheses.[21] Since it does not adequately sample the domain, inferences drawn from the test score (most importantly, that those who fail lack adequate content knowledge to be good teachers) are invalid.

Criterion Validity. Test developers often rely on other tests to demonstrate criterion-related or construct-related validity. For example, Mitchell demonstrated the predictive validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis by correlating scores on those tests with scores on the Stanford Achievement Test. However, she failed to explain what the Stanford Achievement Test measured and how validly it did so.[22]

---

[20] L. Darling-Hammond. "Teaching Knowledge: How Do We Test It?" *American Educator* (Fall 1986), 88.

[21] B. Horner and J. Sammons. *The Test That Fails: An Analysis of the National Teachers Examination in New York* (New York: NYPIRG, 1987), 4–6.

Another approach to demonstrating criterion-related validity relies upon comparisons of test scores with teachers' grades. This, however, undermines a major selling point of standardized tests—that they are an objective substitute for overly subjective teacher judgments.[23] The question is whether the test is more valid than teachers' judgments or some other plausible measure of ability or achievement. The answer is important because test makers will argue that even with low validity, tests can improve decisionmaking as compared with pure chance. However, teacher judgments and other high-quality alternatives are not decisions equivalent to pure chance.[24]

Validity, like reliability, can be measured by statistical methods, which produce numbers called validity coefficients. For many standardized multiple-choice tests, validity coefficients can be quite low, and even high coefficients can result in significant margins of error. School "readiness" tests administered to 4- and 5-year-olds are one example: "Although various readiness tests are correlated with later school performance, predictive validities for all available tests are low enough that 30 to 50 percent or more of children said to be unready [for first grade] will be falsely identified."[25]

No test predicts more than a small fraction of later performance. Employment tests such as the General Aptitude Test Battery (GATB), for example, typically correlate with later performance at the .2–.3 level, meaning they "explain" less than 10 percent of the perceived variance in employee performance.[26]

Inevitably, other indicators exist that also predict some portion of later performance. For example, both high school grades and the SAT predict first-year college performance to some degree. According to the College Board, the statistically weighted, optimum predictive validity coefficient of the SAT correlates at .42 with freshman college performance (thus "explaining" less than 20 percent of the variance in student performance). However, if the SAT score is added to the high school grades (which are stronger predictors), the additional contribution made by the test is a quite low .07.[27] Clearly, the test mostly measures the same area as high school grades, only not as well. Under what circumstances, then, is it reasonable to require an applicant take the test or for a college to use the test results? Bowdoin College found that the minimally increased predictability attained through test scores was more than offset by such negative effects of testing as reducing the diversity of the student body. Similarly, the Harvard Business School found that the Graduate Management Admission Test contributed so little to the admissions process that they now refuse to consider test scores in admissions decisions.[28]

---

[22] B.C. Mitchell. "Predictive Validity of the Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness Analysis for White and for Negro Pupils," *Educational and Psychological Measurement* (1967), 1047–1054. See also, P.H. Johnston, "Assessment in Reading," in P.D. Pearson (ed.), *Handbook of Reading Research* (New York: Longman, 1984), 162. The tendency is for test maker's evidence on criterion-related validity to take the form of "Test A is valid because test B is valid because test C is valid, etc."

[23] Congressional Budget Office. *Educational Achievement: Explanations and Implications of Recent Trends.* (Washington, D.C., Government Printing Office, August 1987).

[24] P. Johnston (1984).

[25] Shepard & Smith (1988). See also, "Mass Academic Testing of Young Children Should Stop, Groups Argue," *Education Week* (Mar. 25, 1988), 5.

[26] R. Seymour. "Why Plaintiffs' Counsel Challenge Tests, and How They Can Successfully Challenge the Theory of 'Validity Generalization,'" *Journal of Vocational Behavior,* vol. 33 (1988), 331–64.

[27] *1988-89 ATP Guide* (1988), 29. J. Crouse and D. Trusheim argue that the contribution the SAT makes above and beyond grades is lower than that reported by the College Board. *The Case Against the SAT* (Chicago: University of Chicago Press, 1988).

It must also be asked, given the limited predictive range of tests, whether other attributes might, for some or all populations, better predict employment success. In recent years, for example, the Federal Government has replaced the Professional and Administrative Career Examination (PACE) with the Individual Achievement Record (IAR), a biographical summary and analysis. IAR scores correlated well with job performance, and the score gap between blacks and whites on the IAR was significantly lower than the gap on the PACE. For applicants with sufficiently high college grade point averages, the Government has concluded that even the IAR is unnecessary.[29]

As validity is not found in the instrument but rather in its use, establishing the validity of any test requires a school, program, business, or government to consider the degree of predictability of the test for different populations of students or applicants. Even if this is done (and typically it is not), two problems remain: The school, program, or job itself may change, and the predictive test may have been used so as to create a self-fulfilling prophecy.

For example, consider athletes who have SAT scores under 700 or ACT scores under 15. Under the National Collegiate Athletic Association (NCAA) regulation Proposition 48, those students may be admitted to a university and receive scholarships if they have adequate grades, but they may not participate in varsity sports during their first year. If their first year grades are adequate, they then can participate in sophomore and subsequent years. NCAA Prop. 42, if ultimately implemented, will alter the policy so that those scoring below the cutoff cannot receive scholarships, regardless of their grades.[30] Prop. 42 clearly implies a prediction: those who score below the cut cannot do college level work. (This is, of course, similar to the claim by the makers of the NTE, noted above, that those who do not pass the NTE could not be good teachers—a claim never proven by predictive validity research.)

A study of University of Michigan athletes showed that, for students who would have been denied scholarships (or entry) for low test scores, 86 percent succeeded as freshman students. That is, the prediction was correct only 14 percent of the time.[31]

However, it may be that many low-scoring athletes need and receive additional academic help. In this scenario, academic assistance could change the context and render the test-score-based prediction false. This raises many value questions: Should the extra help be provided? If so, should it be provided to other low scorers, not just athletes? Can we tell which low scorers would benefit from the extra help?

---

[28] A. Allina. *Beyond Standardized Testing*, 2nd ed. (Cambridge, Mass.: FairTest, 1989).

[29] S. Landers. "PACE To Be Replaced With Biographical Test," *APA Monitor* (April 1989).

[30] *The NCAA News* (Jan. 18, 1989), 1.

[31] T. Walter, *et al.* "Predicting the Academic Success of College Athletes," *Research Quarterly for Exercise and Sport*, vol. 58, no. 2 (1987) 273–79.

Finally, consider the effect of school tracking on the basis of test scores. Typically, those who do not do well on a test are placed in slower tracks. Too often, children from low-income or minority-group backgrounds are the ones who test poorly and then are tracked into low-performance groups. Once placed, they rarely rise to a higher track, in part because the curriculum to which they are exposed is less rich than that in higher tracks.[32] Thus, the initial test that predicted low achievement is proven correct by a self-fulfilling prophecy. If they were not tracked, would the test retain its predictive power? The very existence of "effective schools," schools that succeed with the sorts of students who do not succeed in most school settings, suggests that it would not.[33]

Thus, the use of tests as predictors cannot be divorced from the contexts in which the tests are used. Those contexts may change, and those contexts may be shaped by the tests.

The limitations of predictive criterion validity reinforce the conclusion that no test should be used as a sole or primary criterion for educational or employment decisions.

Construct Validity. Serious doubts also have been raised regarding the general construct-related validity of standardized educational testing. Many test developers do not go beyond content-related validity studies.[34] For example, the widely used and highly respected Iowa Test of Basic Skills "is somewhat lacking when it moves beyond content validity into other validity realms."[35] Professional reviewers of other standardized tests often reach similar conclusions.[36]

Often a test will purport to measure one thing when, in fact, it measures another. Deborah Meier, Principal of Central Park East Secondary School in Manhattan, argues that reading tests do not measure reading but rather measure "reading skills," which is not the same thing.[37] That is, the tests are based on a faulty understanding of reading and learning to read. This is true not only for testing individuals, but also for assessing programs. As Airasian and Madaus write, "Are traditional standardized achievement tests construct valid in terms of inferences about school or program effectiveness? In general, the answer is no."[38] The lack of construct validity has a direct impact on teaching when curriculum becomes dominated by testing.

---

[32] J. Oakes. *Keeping Track: How Schools Structure Inequality* (New Haven: Yale University Press, 1985). See also, L. Shepard and M.L. Smith, *Flunking Grades: Research and Policies on Retention* (Philadelphia: Falmer, 1989).

[33] R. Edmonds. "Effective Schools for the Urban Poor," *Educational Leadership* (October 1979), 15–24.

[34] Madaus & Pullin (1987).

[35] P.W. Airasian. "Review of Iowa Tests of Basic Skills," *Ninth Mental Measurement Yearbook* (1985), 719.

[36] *Ninth Mental Measurement Yearbook* (1985), see reviews of tests listed in note 10.

[37] D. Meier. "Why Reading Tests Don't Test Reading," *Dissent* (Winter 1982–83). See also, A. Bussis, "'Burn It at the Casket': Research, Reading Instruction, and Children's Learning of the First R," *Phi Delta Kappan* (December 1982); C. Edelsky and S. Harman, "One More Critique of Reading Tests—With Two Differences," *English Education* (October 1988).

[38] P. Airasian and G. Madaus. "Linking Testing and Instruction: Policy Issues," *Journal of Educational Measurement* (Summer 1983), 106.

Although many educational tests assume that the underlying trait being measured develops in a relatively consistent fashion among all individuals, developmental researchers generally agree that this is not true.[39] As our knowledge of thinking, learning, teaching, and child development has grown over recent years, standardized tests have not. The WISC-R IQ test, for example, "has remained virtually unchanged since its inception in 1949. . . . Developments in the fields of cognitive psychology and neuroscience have revolutionized our thinking about thinking, but the WISC-R remains the same."[40] The ability of standardized tests to validly measure growth and change in students' knowledge, abilities, or skills is seriously limited by inaccurate views of child development and human learning.

In the work of leading psychometric theoreticians, construct validity has become the essential core of validity, subsuming content and criterion validity.[41] In large part, this is because we enter the realm of underlying hypotheses, theories, and assumptions once we begin to ask questions about the meaning of the content or the effects of the prediction. Tests are not constructed and used independent of theories of knowledge, ability, and performance, as well as theories about the domain to be measured. (For example, the domain of history must be conceptualized to provide a construct that can be measured.) The relationships among theories, tests and test use should be examined as part of construct validity studies. Typically, as indicated above, either the constructs are not considered at all or they are woefully inadequate or outdated.

Messick, among others, has argued that the validity of a test cannot be considered outside of social or educational values or the consequences of its use.[42] This expansion of the concept of construct validity opens up the entire enterprise of testing to serious problems. *If the general social results of testing are harmful, then testing must, in its own terms, be rejected as lacking in validity.*[43]

Consequences of Testing

Ample evidence exists of the effects of testing that are harmful to individuals, to education, and to society as a whole. Individuals are often subjected to educational deprivation or are excluded from admissions, certification, or employment based on test scores. Schooling can be reduced to test coaching through instruction driven by invalid standardized tests. In turn, these become civil rights issues because the negative effects of testing fall most heavily and systematically on those who are most vulnerable and historically victimized: racial minorities and people from low socioeconomic status (SES) backgrounds. Society as a whole then must live with the consequences of the unjust exclusion of many and a damaged educational system.

---

[39] "NAEYC Position Statement . . ." (1988).

[40] J.C. Witt & F.M. Gresham. "Review of WISC-R," *Ninth Mental Measurement Yearbook* (1985), 1716.

[41] Messick (1989, 1988); Cronbach (1988).

[42] Messick (1989, 1988). See also: Cronbach (1988); C.K. Tittle. "Validity: Whose Construction Is It in the Teaching and Learning Context?" *Educational Measurement* (Spring 1989), 5–13; R.E. Schutz. "Faces of Validity of Educational Tests," *Educational Evaluation and Policy Analysis* (Summer 1985), 139–42.

[43] Johnston, for example, argues that the philosophy of science underlying the concept of validity presumes a model of education in which the student and the teacher are both objects. This model, he charges, disempowers student and teacher, with detrimental effects to both as well as to education and society. What is needed, he concludes, is a different conception of science connected to a fundamentally different educational practice—different values and different consequences. (P. Johnston. "Constructive Evaluation and the Improvement of Teaching and Learning," *Teachers College Record* (Summer 1989), 509–28.)

Bias in Testing

Test makers claim that the lower test scores of racial and ethnic minorities and of low-income students simply reflect the biases and inequities that exist in American schools and society. While these problems certainly exist, standardized tests do not just reflect their impact, they compound them.

The use of standardized tests is often defended on the grounds of their "objectivity." But all "objective" really means is that the test can be scored without human subjectivity, by machines.[44] Bias can still creep into the questions themselves. In fact, the purported objectivity of tests is often no more than the standardization of bias.

Researchers have identified several characteristics of standardized tests which could bias results against minority and low-income students and job applicants. Each reflects a focus on the middle to upper class language, culture, or learning style which typifies these exams. As a result, test scores are as much a measure of race/ethnicity or income as they are of achievement, ability, or skill.[45]

To communicate their level of achievement, ability, or skill, test takers must understand the language of the test. Obviously, tests written in English cannot effectively assess those who primarily speak Spanish or some other language and for whom English is a second, partially learned language.[46]

Researchers also have discovered that use of the elaborated, stylized English that is common on standardized exams prevents tests from accurately measuring students who use nonstandard English dialects. These include speakers of Afro-American, Hispanic, Southern, Appalachian, and working-class dialects.[47]

A related type of bias stems from stylistic or interpretive language differences related to culture, income, or gender. For instance, the word "environment" is often associated by Afro-Americans with terms such as "home" or "people" while whites tend to associate it with "air,"

---

[44] B. Hoffman. *The Tyranny of Testing*, (New York: Crowell-Collier: 1962), 60–61.

[45] Some of these characteristics could also lead to gender bias in standardized tests. However, gender bias affects both males and females. Among very young children, some tests appear to be biased against boys ("NAEYC Position Statement . . . " 1988). On the other hand, among older children and adolescents, most bias affects girls (P. Rosser, *Sex Bias in College Admissions Tests*, 3rd edition, Cambridge: FairTest, 1989).

[46] National Coalition of Advocates for Students. *New Voices: Immigrant Students in U.S. Public Schools* (Boston: NCAS, 1988).

[47] M.R. Hoover, R.L. Politzer & O. Taylor. "Bias in Reading Tests for Black Language Speakers: A Sociolinguistic Perspective," *Negro Educational Review* (April–July 1987), 81–98.

"clean," or "earth". Neither usage is wrong. However, on a standardized test only one of these two usages, generally the one reflecting the white usage, will be acceptable.[48]

Similarly, researchers have discovered that individuals exhibit "different ways of knowing and problem-solving" which reflect different styles, not different abilities. These differences are often correlated with race/ethnicity, income level, and gender. Yet standardized tests assume that all individuals perceive information and solve problems in the same way.[49]

Another source of bias appears in questions which assume a cultural experience and perspective which not all test takers share. The WISC-R IQ test, for example, asks "What are you supposed to do if you find someone's wallet or pocketbook in a store?" Children receive two points for answering "Give it to the store owner," one point for answering "Look to see who it belongs to," and no points for replying, "Make believe you didn't see it . . . Don't keep it."[50] Yet youngsters living in high crime neighborhoods may choose to ignore the wallet or pocketbook for fear they would be accused of stealing it. Researchers at Johns Hopkins found that inner-city black children often answered WISC-R questions "incorrectly" for a variety of reasons other than lack of knowledge or ability.[51] Giving the wrong answer to just a few such questions can cause one's "IQ" (or "achievement") to appear sharply lower, with possibly life-scarring results.

Ironically, even efforts to decontextualize test content has been shown to work against minority and low-income youths. Middle-class whites are more apt to be trained through cultural immersion to respond to questions removed from context and to repeat information the test taker knows the questioner already possesses. Heath found that working-class black children, in their communities, were rarely asked questions to which the questioner already knew the answer, like those found on standardized tests.[52]

Students also tend to perform better on tests when they identify with the subjects of the test questions. Research on Mexican Americans, African Americans, and females all reveal that "items with content reference of special interest" to each group seem to improve their test scores.[53] Unfortunately, questions on standardized tests remain disproportionately about and for upper income white males.

---

[48] J. Loewen. "Possible Causes of Lower Black Scores on Aptitude Tests" (unpublished research report, 1980).

[49] O. Taylor and D.L. Lee. "Standardized Tests and African Americans: Communication and Language Issues," *The Negro Educational Review* (April-July, 1987), 67–80.

[50] D. Wechsler. *Wechsler Intelligence Scale for Children–Revised* (New York: The Psychological Corporation, 1974), 176.

[51] J. Butler. "Looking Backward: Intelligence and Testing in the Year 2000," *National Elementary Principal* (March/April 1975), 73–74.

[52] T. Meier. "The Case Against Standardized Achievement Tests," *Rethinking Schools* (vol. 3, no. 2, 1989), 12. See also Levidow (1987). S.B. Heath. *Ways With Words: Language, Life, and Work in Communities and Classrooms* (New York: Cambridge University Press, 1983), cited in Meier (1989).

The timed format of many tests also can be a source of bias. Several studies have found that speededness is a factor for lower scores of blacks, Hispanics, and women.[54]

These and other forms of bias are reinforced by the procedures used to construct and norm tests. For example, questions that might favor minorities are apt to be excluded for not fitting the "required" statistical properties of the test. Even if minorities are included in the test companies' samples in accord with their portion of the overall population, at least three quarters of the sample will be white. Moreover, African Americans, American Indians, and Hispanics are disproportionately among the low-scoring group. In general, test makers discard those questions on which low scorers do well but high scorers do poorly.[55] As a result, a sample question on which blacks do particularly well but whites do not is likely to be discarded for the compound reason that blacks are a minority and generally score low.

Nonetheless, test companies maintain that they effectively screen out biased questions. Though they subject items to review by experts who supposedly can detect bias, such screening is of low reliability.[56] Though most major test makers also apply some form of statistical procedure, even when bias is found items are not necessarily removed. Moreover, the procedures themselves are often problematic. Typically, they presume the independence of the part (the item) from the whole; but if the entire test is biased in form or content, item analysis will not reveal it.[57]

---

[53] A.P. Schmitt and N.J. Dorans. "Differential Item Functioning for Minority Examinees on the SAT," (Paper for American Psychological Association annual meeting, 1987). For research on Hispanics, see A.P. Schmitt, "Unexpected Differential Item Performance of Hispanic Examinees on the SAT-Verbal, Forms 3FSAO8 and 3GSAO8," (unpublished statistical report of the Educational Testing Service, 1986). Dr. Schmitt concluded that Mexican American students scored significantly higher than expected on a reading comprehension passage concerned with lifestyle changes in Mexican American families. For research on blacks, see Hoover, Politzer & Taylor (1987), who report that Dr. Darlene Williams found "the use of pictures showing Blacks and related to Black culture raised IQ scores for all Black children." For research on females, see J.W. Loewen, P. Rosser & J. Katzman, "Gender Bias in SAT Items," (Paper presented at the AERA Annual Convention, New Orleans, La., Apr. 5, 1988). Also, the mathematics section of the WISC-R test includes eight questions about 13 boys or men who save money on purchases, trade fairly, cleverly divide their efforts and money and work at jobs, compared to only one question featuring a girl who loses her hair ribbon (Wechsler, 1974).

[54] N.J. Dorans, A.P. Schmitt, W.E. Curley. "Differential Speededness: Some Items Have DIF Because of Where They Are, Not What They Are" (paper for the National Council on Measurement in Education annual meeting, 1988.) G.I. Maeroff. "Reading Test Time Limits Are Criticized," *New York Times* (Jan. 19, 1985). See also Schmitt and Dorans (1987). P. Rosser. *The SAT Gender Gap: Identifying the Causes* (Washington, D.C.: Center for Women Policy Studies, 1989).

[55] B. Hoffman (1962), pp. 54-56.

[56] J.D. Scheuneman. "A Posteriori Analyses of Biased Items," in R.A. Berk, *Handbook of Methods for Detecting Test Bias* (Baltimore: Johns Hopkins, 1982). L.A.Shepard. "Identifying Bias in Test Items," in B.F. Green, *New Directions for Testing and Measurement: Issues in Testing—Coaching, Disclosure and Ethnic Bias*, no. 11. (San Francisco: Jossey-Bass, 1981).

[57] Shepard (1981). Berk, ed., (1982), chap. 9, "Methods Used by Test Publishers to 'Debias' Standardized Tests."

BEST COPY AVAILABLE

The impact of bias in testing is that test scores underestimate the abilities of minority and low income students and applicants. This was demonstrated by an experiment in which two alternate forms of the NTE General Knowledge test, containing content less likely to be unfamiliar to blacks but otherwise possessing similar properties, were constructed and tested by ETS researchers. On one alternative test, black examinees performed better than whites. On the second, they did less well but better than on the traditional NTE, on which the black pass rate tends to be one-half that of the white rate.[58]

The NTE and similar tests deserve particular attention for the effects they have on the minority teaching force. While the minority student population in the U.S. will exceed one third of the total by the year 2000, only 5 percent of the teaching force will be minority if current trends prevail. More than half the African American and Hispanic applicants fail teacher tests, which lack content and predictive validity, making testing a major factor in the reduction of the minority teaching force.[59] The absence of minority teachers causes harm not only to minority students, who lose role models and teachers who understand their cultural background, but also to majority students, who lose the opportunity to be exposed to minority adults in positions of responsibility.

Bias can render a test invalid for the groups against which it discriminates. But the same factors also weaken test validity for those who benefit from the bias. For example, men from all ethnic groups and income levels score higher on the SAT than do women from comparable groups, though women earn higher grades in both high school and college. This bias lowers the test's validity for both groups by overpredicting men's grades and underpredicting women's. However, damage from the sex bias falls solely on women who, because of their lower scores, may be denied admission or scholarships or suffer a loss of personal and social esteem.[60]

By ignoring the skills, abilities, life experiences, learning styles, languages, and cultures of minority and low-income groups, testing devalues those people and their attributes. In education, this encourages a pedagogy based on correcting deficits, not one based on building from strengths. In education and employment, this perpetuates a "requirement" that only "white" styles are acceptable.

---

[58] D.M. Medley and T.J. Quirk. "The Application of a Factorial Design to the Study of Cultural Bias in General Culture Items on the National Teacher Examination." *Journal of Educational Measurement* (vol. II, no. 4, Winter 1974). See also, R.K. Hackett *et al.* "Test Construction Manipulating Score Differences Between Black and White Examinees: Properties of the Resulting Tests" (Princeton, N.J.: Educational Testing Service, 1987).

G.P. Smith. *The Effects of Competency Testing on the Supply of Minority Teachers: A Report Prepared for the National Education Association and the Council of Chief State School Officers*, (University of North Florida, Jacksonville: 1987).

[59] G.P. Smith (1987).

[60] P. Rosser (1989).

Impact of Testing on Schooling

Historically, standardized tests were one of several educational tools used to assess student achievement and to diagnose academic strengths and weaknesses. In recent years, however, standardized tests have become not only the primary criterion used by many schools for making decisions affecting students, but also major forces in shaping instruction and assessing the quality of teaching and the schools.

Impact on Student Progress. By controlling or compelling student placement in various educational programs, standardized tests perpetuate and even exacerbate existing inequities in educational services, particularly for minority and low-income students.

One clear example is tracking, which has been shown to harm low-track students without necessarily helping those in higher tracks do better than they would in heterogenous groupings. In large part this is because those with low test scores are presumed unable to master complex material and are fed a "dumbed-down" curriculum.[61]

Standardized test results also lead to larger numbers of racial and ethnic minorities being placed in special education and remedial education programs. Blacks, for example, are two to three times as likely to be in classes for the educable mentally retarded as are whites.[62]

Standardized tests also perpetuate the domination of white upper middle-class students in "advanced" classes. In New York City, IQ tests are used in some districts to place children in "gifted and talented" programs, creating white, upper middle-class enclaves in districts whose enrollment is dominated by racial and ethnic minorities.[63] Overall, test use both narrows the educational opportunities available to many segments of our student population and maintains the isolation of racial and social groups and classes. At the same time, standardized tests, particularly when used as promotional gates, can act as powerful exclusionary devices. Research has demonstrated that, for a student who has repeated a grade, the probability of dropping out prior to graduation increases by 20 to 40 percent.[64] Thus, students who are not promoted because they fail an often unreliable, invalid, and biased standardized test are more likely to drop out of school.

The impact of standardized tests is particularly devastating when used to determine "readiness" for kindergarten or first grade. As noted above, these tests are among the least valid and reliable and are among the most difficult to administer under relatively uniform conditions.

---

[61] J. Oakes (1985). See also the 1988 NAEP reports on Reading and Math (ETS, Princeton) for the types of instruction offered in low tracks.

[62] J.D. Finn. "Patterns in Special Education Placement as Revealed by the OCR Surveys," in K. Heller, W. Holtzman, S. Messick, eds., *Placing Children in Special Education* (Washington, D.C.: National Academy Press, 1982).

[63] A. Cook, Community Studies, Inc., New York City, N.Y., (Personal communication, April 1988).

[64] Massachusetts Advocacy Center. "Memorandum to the Boston School Committee" (June 19, 1987) quoting from Office of Educational Assessment, New York City Board of Education, "Evaluation Update on the Effect of the Promotional Policy Program" (Nov. 12, 1986). See also, M.L. Smith and L.A. Shepard.,"What Doesn't Work: Explaining Policies of Retention in the Early Grades," *Phi Delta Kappan* (October 1987), 129–34.

Moreover, Shepard and Smith, after examining 14 controlled studies on the effects of kindergarten retention, concluded that retention provided no increase in subsequent academic achievement while imposing a significant social stigma on the retained students.[65]

Nor does the use of standardized tests affect only low-achieving students. High-achieving students or those whose interests stray from the basics are likely to be frustrated by a narrowed curriculum, which has been "dumbed-down" in response to standardized exams, particularly minimum competency tests. These students too are likely to drop out in higher numbers.[66]

Impact on Educational Goals and Curriculum. Paul LeMahieu and Richard Wallace of the Pittsburgh schools note the inevitability of testing's impact on schooling: "It is untenable to agree that achievement is the product, and that test scores are its measure, and then assert, 'Please don't pay too much attention to the scores.'"[67] The result of the emphasis on testing is, as George Madaus observed, that rather than being "compliant servants," tests have become "dictatorial masters."[68]

Children go to school not just to learn basic academic skills, but also to develop the personal, intellectual and social skills to become happy, productive members of a democratic society. Unfortunately, the current emphasis on standardized tests threatens to undermine this educational diversity by forcing schools and teachers to focus on narrow, quantifiable skills at the expense of more complex, academic and nonacademic abilities.

The narrowing of diversity is particularly true for young children. As the National Association for the Education of Young Children (NAEYC) recently cautioned: "Many of the important skills that children need to acquire in early childhood—self-esteem, social competence, desire to learn, self-discipline--are not easily measured by standardized tests. As a result, social, emotional, moral, and physical development and learning are virtually ignored or given minor importance in schools with mandated testing programs."[69]

Many schools have embarked on a single-minded quest for higher test scores even though this severely narrows their curriculum.[70] For example, Deborah Meier noted that students read "dozens of little paragraphs about which they then answer multiple-choice questions"—an approach that duplicates the form of the tests the students take in the spring.[71] And Gerald Bracey, former Director of Research, Evaluation and Testing in the Virginia Department of Education, observed that some students were not taught how to add and subtract fractions because the State's minimum competency test included questions on multiplication and division of fractions, but not on their addition and subtraction.[72]

---

[65] Shepard & Smith (1988), 34. See also, Smith & Shepard (1987), and Shepard and Smith, eds. (1989).

[66] "Student Competency Exams Present Major Barrier to Minority Students," *Education Daily* (Aug. 27, 1987), 3.

[67] P.G. LeMahieu and R. C. Wallace, Jr. "Up Against the Wall: Psychometrics Meets Practice," *Educational Measurement* (Spring 1986), 12–16.

[68] G.F. Madaus. "The Influence of Testing on the Curriculum," in *87th Yearbook of the National Society for the Study of Education*, Part I (1988), 83–121. Since this was written, evidence of the disastrous effects of testing on curriculum and instruction has mushroomed. See the papers presented at the special American Educational Research Association conference on national testing, *Phi Delta Kappan* (in press, November 1991).

[69] National Association for the Education of Young Children. *Testing of Young Children: Concerns and Cautions* (Washington, D.C.: NAEYC, 1988). See also, "NAEYC Position Statement . . . " (1988).

[70] G.F. Madaus (1986). See also, H.C. Rudman. "Testing Beyond Minimums," *ASAP Notes* (Occasional Paper No. 5, 1985), 1–36.

Sometimes, the curriculum is narrowed simply because "testing takes time, and preparing students for testing takes even more time. And all this time is time taken away from real teaching."[73]

Unfortunately, a closer link between tests and curriculum has become a very conscious goal for some educators. School systems in at least 13 States and the District of Columbia are seeking to "align" their curriculum so that students do *not* spend hours studying materials upon which they will never be tested regardless of the value or benefits which could be derived from that effort.[74] Curriculum alignment "subordinates the process of curriculum development to external testing priorities, namely the State minimum-competency exam. Thus, the curriculum falls in line with the test, and, for all intents and purposes, the test becomes the curriculum."[75]

The educational price paid for allowing tests to dictate the curriculum can be high. Julia R. Palmer, Executive Director of the American Reading Council, recently wrote, "[T]he major barrier to teaching reading in a commonsense and pleasurable way is the nationally normed standardized second grade reading test." Ms. Palmer explains that the test questions force teachers and students to focus on "reading readiness" exercises and workbooks in their early grades and not on reading. As a result, many students become disenchanted with reading because they rarely get a chance to participate in it or to read anything of real interest to them.[76]

Mathematics instruction has also been harmed by the emphasis on testing. Constance Kamii reports that the tests are unable to distinguish between students who understand underlying math concepts and those who are only able to perform procedures by rote and are thus unable to apply them to new situations. Teaching to the test, therefore, precludes teaching so that children grasp the deeper logic.[77] The National Council of Teachers of Mathematics has concluded that unless assessment is changed, the teaching of math cannot improve.[78]

---

[71] G.F. Madaus. "Test Scores as Administrative Mechanisms in Educational Policy," *Phi Delta Kappan* (May 1985), 616.

[72] "Some 'Teach' to the Test," *The [Newport News, VA] Daily Press* (June 15, 1987), C1.

[73] A.E. Wise. "Legislated Learning Revisited," *Phi Delta Kappan* (January 1988), 330. D.W. Dorr-Bremme & J.L. Herman. *Assessing Student Achievement: A Profile of Classroom Practices* (Los Angeles: Center for the Study of Evaluation, UCLA, 1986).

[74] L. Olson. "Districts Turn to Nonprofit Group for Help in 'Realigning' Curricula to Parallel Tests," *Education Week* (Oct. 18, 1987), 1 & 19.

[75] P.S. Hlebowitsch. Letter to the editor, *Education Week* (Nov. 18, 1987), 21.

[76] J.R. Palmer. Letter to the editor, *New York Times* (Dec. 14, 1987). See also, J.T. Guthrie, *Indicators of Reading Education* (Center for Policy Research in Education: New Brunswick, N.J., 1988), which concludes that the strengthening of students' reading skills goes hand-in-hand with finding better ways to measure reading achievement. The primary shortcoming of reading tests is that they don't reflect the complexity of the reading process. See also, Edelsky and Harman (1988).

Just as curriculum has been narrowed, so too have textbooks. Diane Ravitch argues that "textbooks full of good literature began to disappear from American classrooms in the 1920s, when standardized tests were introduced. Appreciation of good literature gave way to emphasis on the 'mechanics' of reading."[79] Similarly, a report by the Council for Basic Education concluded that the emphasis on standardized tests and curriculum alignment were among the main causes of the increasingly poor quality of textbooks. The report noted that "instead of designing a book from the standpoint of its subject or its capacity to capture the children's imagination, editors are increasingly organizing elementary reading series around the content and time of standardized tests . . . As a result, much of what is in the textbooks is incomprehensible."[80]

The narrowing of curriculum is a virtually unavoidable byproduct of emphasizing instruments of limited construct validity that utilize a multiple-choice format. Not only do reading tests not test reading and math tests not test math, but the format dictates against them ever being able to measure the essential content or construct. As teaching becomes test coaching, real learning and real thinking are crowded out in too many schools.

Among the instructional casualties are efforts to improve what is now labeled "higher order thinking skills." Standardized tests, including many required under State school reform laws, focus on basic skills, not critical thinking, reasoning or problem solving. They emphasize the quick recognition of isolated facts, not the more profound integration of information and generation of ideas.[81] As Linda Darling-Hammond of the Rand Corporation concluded, "It's testing for the TV generation—superficial and passive. We don't ask if students can synthesize information, solve problems, or think independently. We measure what they can recognize."[82]

Moreover, several studies have demonstrated that "teaching behaviors that are effective in raising scores on tests of lower level cognitive skills are nearly the opposite of those behaviors that are effective in developing complex cognitive learning, problem-solving ability, and creativity."[83] Because children learn "higher skills" (the integration, use, and creation of knowledge) from the very start, it is not always necessary to teach basic skills and higher skills sequentially. Indeed the very process of learning is itself an active, "higher order" process, negating the artificial distinction between learning "basic" and "higher order" skills.[84] Testing stands in the way of needed curricular change.

---

[77] C. Kamii. *Young Children Continue to Reinvent Arithmetic, 2nd Grade* (New York: Teachers College Press, 1989).

[78] National Council of Teachers of Mathematics. *Curriculum and Evaluation Standards for School Mathematics* (1989). See also, National Research Council of the National Academy of Sciences, *Everybody Counts—A Report to the Nation on the Future of Mathematics Education* (Washington, D.C.: National Academy Press, 1989).

[79] E.B. Fiske. "America's Test Mania," *New York Times* (Apr. 10, 1988), Section 12, p. 20.

[80] H. Tyson-Bernstein. *A Conspiracy of Good Intentions: America's Textbook Fiasco* (Washington, D.C.: Council for Basic Education, 1988). See also, K.I. Goodman, *et al.*, *Report Card on the Basal Readers* (Katonah, N.Y.: Owen Publishers, 1988).

[81] A. Bastian, *et al.* *Choosing Equality: The Case for Democratic Schooling* (Philadelphia: Temple University Press. 1986), 73. See also, Frederickson (1984).

[82] T. Fiske. *New York Times* (Apr. 10, 1988), 20.

[83] M.C. McClellan. "Testing and Reform," *Phi Delta Kappan* (June 1988), 769.

For students who are tracked into programs for "slow learners" due to their low test scores, rote and basic skills are emphasized. This bores, frustrates, and alienates both students and teachers in a dialectic that fosters student resistance to the schooling that is offered, often in the form of disciplinary problems.[85] The result is a lack of learning. The common solution to the ensuing low test scores is more "basics" and more testing in a program designed to raise the scores. Not surprisingly, many students at best graduate hating school while having learned little, and at worst drop out into the ranks of the chronically unemployed and unemployable. This scenario most typically affects minority and low-income students.

Standardized testing is clearly not the only culprit, but through its effects on texts, pedagogy and goals, it is a major problem. The continuing overemphasis on testing and what can be measured by tests will only make the situation worse and hinder the possibility of solution.

Impact on Local Control. Because standardized tests increasingly determine what is taught and how it is taught, parents and other citizens are losing their traditional control over the public schools. This shift of power from local communities to State and National Government reduces the level of input and influence available to both parents and teachers in the management of the schools. This, in turn, reduces "the responsiveness of schools to their clientele and so reduces the quality of education" available in those schools.[86]

Local control over the schools is also being lost to private organizations, namely the test developers. Despite the significant and growing role their products play in educational decisions, testing manufacturers face little government regulation or supervision. Unlike other businesses, such as communications, food and drugs, transportation, and securities, there are virtually no regulatory structures at either the Federal or State level governing the billion-dollar-a-year testing industry.

States and school districts have neither the expertise nor the resources either to independently develop and validate standardized tests or to adequately investigate claims by test developers regarding test validation.[87] Even if the expertise and resources did exist, the secrecy which is rampant in the testing industry would likely prevent any effective outside evaluation. As the late Oscar K. Buros, editor of the *Mental Measurement Yearbook*, lamented, "It is practically impossible for a competent test technician or test consumer to make a thorough appraisal of the construction, validation, and use of standardized tests . . . because of the limited amount of trustworthy information supplied by the test publishers."[88]

---

[84] "NAEYC Position Statement . . . " (1988). See also G. Bracey, "Advocates of Basic Skills 'Know What Ain't So'," *Education Week* (Apr. 5, 1989), 32; Resnick, L. and D.P. Resnick. "Assessing the Thinking Curriculum: New Tools for Educational Reform," in B.R. Gifford and M.C. O'Connor, eds., *Future Assessments: Changing Views of Aptitude, Achievement and Instruction* (Boston: Kluwer Academic Publishers, 1989).

[85] The seminal work on this aspect of resistance is P. Willis, *Learning to Labor*, (Lexington, Mass.: Heath, 1977). See also H. Giroux, "Theories of Reproduction and Resistance in the Sociology of Education: A Critical Analysis," *Harvard Educational Review*, vol. 53, no. 3. The "middle class" cultural basis of the school is also opposed by those from other class, race or cultural backgrounds.

[86] A.E. Wise. "Legislated Learning Revisited," *Phi Delta Kappan* (January 1988), 328–333. See also, A. Porter. "Indicators: Objective Data or Political Tool?," *Phi Delta Kappan* (March 1988), 503–508.

[87] Madaus & Pullin (1987), 3-4.

Testing: An Invalid Enterprise

In sum, current standardized, multiple-choice tests are severely flawed instruments. Their overuse and misuse cause substantial individual and social harm. Many factors contribute to these problems:

- Test makers make assumptions about human ability that cannot be proven.
- No test is sufficiently reliable to be used as sole or primary criteria.for decisionmaking.
- The content validity of tests is inadequate because they cannot measure the complex material contained in most learning or performance domains.
- Predictive criterion validity is too low to use tests as sole or primary criteria for decision making. The limited degree of validity that does exist often results from self-fulfilling prophecies.
- The construct validity of tests is likewise inadequate: tests often do not measure the traits they claim to measure or do so only poorly.
- Standardized exams often fail to accurately measure persons from different backgrounds, and test results are used to segregate and devalue persons from minority groups.
- The effects of testing not only cause irreparable harm to many individuals, they also are destructive to the educational process as a whole. It is low-income and minority-group students who are most often subjected to the poorest, narrowest, most rigidly test-driven curriculum and instruction.
- Tests contribute to the exclusion of minorities from colleges and universities and are major roadblocks to equal opportunity employment in the U.S.

If, as some of testing's foremost theoreticians suggest, the validity of testing is inseparable from its social consequences, then standardized, multiple-choice testing is substantially invalid.

In education, testing is, at best, hopelessly inadequate for promoting necessary school reform. At worst, overreliance on testing will preclude reform. In either case, the continued domination of testing means that millions of students, predominantly those most in need of improved education, will be dumped into dead end tracks and pushed out of school. To prevent damage and to allow needed reforms, testing must become an occasional adjunct, used for attaining basic but limited information about educational policies.

In university admissions, testing is largely unnecessary. Schools ought not to require students to pay for and take exams that add little to a school's ability to predict success. And in employment, the absence of strong predictive validity coupled with often-weak content validity means that employment decisions should not be made on the basis of test scores and alternatives should be used for selecting and promoting employees.

---

[88] O. Buros. "Fifty Years in Testing: Some Reminiscences, Criticisms, and Suggestions," *Educational Researcher* (July-August, 1977), 14.

Appropriate Assessments

Better methodologies for assessment have been and are being devised to serve the needs of instruction, learning, and evaluation. Most rely on some form of what Gardner refers to as "process and product portfolios."[89]

Instead of indirectly measuring an often ill-defined and unanalyzed construct, alternative assessments can use direct evidence of the trait itself, e.g., writing samples on meaningful topics collected over time, rather than an hour's worth of multiple-choice sentence correction problems. Teacher observations themselves can be recorded and summarized in a systematic manner.[90]

Assessment, properly done, can be of great help to instruction and learning. It can encourage critical thinking and creativity. Teachers can pinpoint not only what a student knows, but how the student best learns. High quality alternative processes would ensure the use of multiple forms of measurement leading to more valid measures of competence, achievement, and ability.

Across the country, in schools, districts, States, and research programs, authentic and appropriate assessments are being designed and implemented. In North Carolina, which banned the use of achievement tests in grades one and two, developmentally appropriate assessments tied to the State's curriculum will be implemented in the fall of 1989. Missouri is similarly constructing developmentally appropriate assessments for young children. The National Association for the Education of Young Children (NAEYC) has convened a working group to develop assessments that can be used across the Nation.[91]

Appropriate assessments are being developed not only for young children. California plans to replace multiple-choice tests for all its California Assessment Program exams over the next 5 years. New York recently experimented with a hands-on grade four science exam. Connecticut is pioneering a variety of alternatives at the high school level. And the Pittsburgh schools are developing authentic assessments for use in a variety of grades and subjects. Arizona, Kentucky, Maryland, and Vermont are committed to making their State assessments primarily performance assessments.[92] A number of colleges have found they can select strong student bodies without using standardized tests.[93] And the replacement of PACE with an alternative process indicates that better methods than tests exist for employment selection.[94]

---

[89] D.A. Archibald & F. M. Newmann. *Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School* (Reston, VA: National Association of Secondary School Principals, 1988). See also the special issue of *Educational Leadership* on "Redirecting Assessment" (April 1989) and articles on the same theme in *Phi Delta Kappan* (May 1989). H. Gardner. "Assessment in Context: The Alternative to Standardized Testing" (Berkeley: Paper for the National Commission on Testing and Public Policy, 1988). Gardner has written a number of other articles on the same topic. Knowledge and experience in this area is growing rapidly. FairTest provides regular updates on practice in its newsletter, the *FairTest Examiner* (Cambridge, Mass.).

[90] Gardner (1988). Johnston (1989). See also sources in note 90.

[91] North Carolina Department of Public Instruction. *Grades 1 and 2 Assessment* (Raleigh, 1989). FairTest. "Missouri Developing Alternatives to Standardized Testing," *FairTest Examiner* (Winter 1989), 7. Personal discussion with S. Bredekamp, NAEYC.

But no matter how well crafted, improved assessment is not a panacea. Alternatives must be carefully designed so as not to reproduce the biases, inaccuracies, or damage to students and curriculum of standardized educational and employment tests. Replacing the bias built into standardized tests with the bias of the individual teacher, school, or employer would not be progress. Thus, alternatives must build in means to detect bias, and where found, procedures to correct it.[95]

The FairTest Agenda for Testing Reform

FairTest's agenda for testing reform reflects its concern over the misuse of standardized tests. Major reforms in the instruments themselves and sharp controls on their use are necessary to make tests fair, accurate, open, and relevant.[96] The FairTest Agenda is guided by four basic principles:

- *Tests must be properly constructed, validated, and administered.* Tests should measure pertinent, not extraneous, knowledge differences among students or applicants. Questions must be relevant to the knowledge, abilities, or skills being tested. Test items and instructions should be written clearly and accurately.

The tests themselves should take into account the diversity of language, experience and perspective embodied in the test-taking population. At the same time, questions and scoring procedures should acknowledge the complexity and diversity of intelligence and individual development.

Test validation should ensure that the content of the test matches the content of what is taught or done on the job. But test developers cannot stop at content validation. They must document assumptions about the relationship between test results and future performance. At the same time, they must demonstrate that test results are accurately related to the underlying knowledge, skills, and abilities the test claims to measure.

---

[92] Personal conversation with R. Mitchell, Council for Basic Education (for California). G. Wiggins of the National Center on Education and the Economy (Rochester, N.Y.) has extensive material on authentic assessments. Both Mitchell and Wiggins are working on books on this topic. For Pittsburgh, see D.P. Wolf, "Portfolio Assessment: Sampling Student Work," *Educational Leadership* (April 1989), 35–40. For State information, see *FairTest Examiner* Summer 1990, Summer 1991.

[93] Allina (1989).

[94] Landers (1989).

[95] Gill and Levidow, eds. (1987), section on "Assessment," 210–267.

[96] Medina and Neill (1988), 24–26. "The FairTest Agenda," *FairTest Examiner* (1987, vol.1, no.3), 16. See also, National Forum on Assessment, *Criteria for Evaluation of Student Assessment Systems* (Washington, D.C., and Boston, Mass.: Council for Basic Education and FairTest, 1991).

- *Tests should be open.* Educators, schools, test takers and independent researchers should all have access to the descriptive and statistical data needed to verify test publishers' claims regarding test construction and validation. This should include the release of questions used on previous tests, as well as data on test results grouped by race/ethnicity, gender, socioeconomic status, geographical residence, and other demographic categories. Users should make public their own procedures for test administration and guidelines for use of test scores.
- *Tests should be viewed in the proper perspective.* Both test developers and test users should work to ensure that test results are properly interpreted and employed by schools, colleges and universities, employers, policymakers, test takers, and the general public. As the 1974 *Standards for Educational and Psychological Tests* states: "A test score should be interpreted as an estimate of performance under a given set of circumstances. It should not be interpreted as some absolute characteristic of the examinee or as something permanent and generalizable to all other circumstances." This standard has too often been ignored by those who use test results. At a minimum, test scores should not be used as the sole or primary factor in educational or employment decisions.

Test developers and test users must recognize that standardized tests are only limited measures of educational reality. Used alone, they present distorted pictures of what they seek to measure and often undermine both educational quality and equal opportunity. Both test developers and test users have the obligation to promote a proper, reasonable, and limited use of standardized tests as one of a series of assessment mechanisms.

- *Appropriate and authentic assessment instruments should be used instead of standardized tests, to the extent possible.* Standardized multiple-choice tests can only measure a very limited range of knowledge, abilities, and skills. Both new technologies and greater understanding of teaching and learning provide opportunities to expand our capability to more fully and accurately measure a greater range of knowledge, abilities, and skills. Educators and employers should invest in developing and using these methods. These can be used to diagnose the strengths and weaknesses of students and workers in order to help them learn, rather than to sort, stratify, or segregate them. More accurate assessments can potentially help both test takers and institutions, though these too must be critically assessed to ensure they do not contribute, as do standardized multiple-choice tests, to inequality.

A Legal and Policy Perspective

By Clint Bolick*
Landmark Legal Foundation Center for Civil Rights

I am pleased to submit this written statement to supplement my verbal testimony before the Commission on June 16, 1989. I represent the Landmark Legal Foundation Center for Civil Rights, a Washington-based law center committed to the advancement of equality under law and fundamental individual rights.

My comments are limited to the legal and policy aspects of testing.[1] I am not a psychologist or a statistician, and therefore take no position on whether tests are intrinsically good or bad, or whether and when people should use them. Those questions, in my view, are generally better committed to the sound judgment of those who use tests, based on the evidence available to them.

The legal limits placed on that discretion, however, have important implications for those who use tests, for those who take tests, and for society as a whole. The legal landscape surrounding the use of tests has recently changed significantly; thus my comments will focus on those changes and their potential effects.[2]

The Center for Civil Rights' interest in the legal aspects of testing is multifaceted, and I would summarize our position on various current issues related to testing as follows:

- We are concerned that tests are often used by State governments and by private entities acting under color of State law as anticompetitive devices to arbitrarily screen out qualified individuals from gaining certification to practice their chosen professions, which in such instances denies the individual's fundamental civil right to pursue a trade or profession free from arbitrary or excessive regulation.
- We are concerned that nondiscriminatory testing devices are wrongfully proscribed pursuant to the misconceived notion that all statistical disparities among races or sexes are the result of discrimination, a notion that leads to racial quotas or the abandonment of tests.
- We are concerned that, as a subset of the second issue, individuals are prohibited in some instances from taking tests solely on account of their race. Our law suit in *Crawford v. Honig*,[3] which I will discuss later in this testimony, illustrates this problem.

---

[*] Since writing this paper, Mr. Bolick has become the vice president and director of litigation of the Institute for Justice, Washington, D.C. Before becoming a director of the Landmark Legal Foundation, Mr. Bolick served as an attorney for the United States Department of Justice, Civil Rights Division (1986-87) and for the United States Equal Employment Opportunity Commission (1985-86). He is author of *Changing Course: Civil Rights at the Crossroads* (New Brunswick, NJ: Transaction Books, 1988).

[1] I have previously addressed these issues in "Legal and Policy Aspects of Testing," 33 *Journal of Vocational Behavior* 320 (1988). The entire issue of the *Journal* was devoted to these issues.

[2] More specifically, my comments will focus primarily on *employment* testing, although the general principles apply to educational testing.

[3] No. C-89-0014-RFP (N.D. Cal.).

Since much of the current controversy focuses on the second issue (the ability of employers or educators to use tests), and since the recent legal developments speak directly to that issue, I will focus most of my attention there.

The Supreme Court's Decisions: Debunking Flawed Conventional Wisdom

Prior to decisions by the United States Supreme Court in its recently completed term, the legal construct employed by many courts led almost automatically to the abandonment or invalidation of tests, regardless of whether they were discriminatory in any real sense. This result was produced by the judicially crafted burdens of proof, which made it relatively easy to challenge tests but nearly impossible to defend them.

This development is contrary to the express intent of Title VII's framers, who made it clear that the law was aimed at eradicating discrimination from the employment market while leaving employer discretion otherwise intact. Senator Hubert Humphrey, the principal architect of Title VII, emphasized that the law "does not limit an employer's freedom to hire, fire, promote, or demote for any reasons—or no reasons—so long as his action is not based on race."[4]

The provisions of Title VII reflect this intent. Section 703(j) of Title VII provides that the law does not require:

> preferential treatment to any individual or group . . . on account of an imbalance which may exist with respect to the total or percentage of persons of any race, color, religion, sex, or national origin employed . . . in any comparison with the total number or percentage in any community . . . or in the available workforce. . . .

Likewise, section 703(h) further provides that it shall not "be an unlawful practice for an employer to give and to act upon the results of any professionally developed ability test provided such test . . . is not designed, intended, or used to discriminate. . . ."

The legislative history and language thus make it clear that Title VII was not intended to require employers to abandon nondiscriminatory employment practices or to seek racially balanced work forces. Indeed, testing devices obviously provide one possible method to *avoid* discrimination since by definition they treat all individuals the same. The goal of Title VII in the testing context, then, is not to enjoin the use of tests generally, or even those that produce racially disproportionate results, but rather to identify and prohibit only those tests that are used as a subterfuge for discrimination.

That is precisely the role the "adverse impact" doctrine, as originally set forth in *Griggs v. Duke Power Co.*,[5] was intended to play. Prior to *Griggs*, the only method by which to prove discrimination in the absence of direct evidence of discriminatory intent was "disparate treatment"—that is, situations in which similarly situated persons of different races are treated differently, which gives rise to a rebuttable presumption that the explanation for the different treatment is discrimination.

---

[4] 110 Cong. Rec. 5423 (1964).

[5] 401 U.S. 424 (1971).

But not all situations are amenable to disparate treatment analysis. *Griggs* presented the question whether an employer's requirement of either a high school diploma or a passing score on a standardized general intelligence test was permissible when "(a) neither standard is shown to be significantly related to successful job performance, (b) both requirements operate to disqualify Negroes at a substantially higher rate than white applicants, and (c) the jobs in question formerly had been filled by white employees as part of a longstanding practice of giving preference to whites."[6]

The Court's answer, not surprisingly, was no: the job requirements, which produced adverse racial impact but did not predict "a reasonable measure of job performance,"[7] the Court concluded, "operate[d] to 'freeze' the status quo of prior discriminatory employment practices."[8] Since "[w]hat is required by Congress is the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate,"[9] the Court ruled the employment requirements invalid under Title VII.

The adverse impact construct is a logical way of ferreting out "covert" instances of discrimination. For example, an all-white community surrounded by black suburbs that adopts a "residency" requirement for municipal jobs is fairly clearly engaging in racial discrimination if it cannot show a business purpose for its requirement.[10]

But this rational application of adverse impact to uncover hidden discriminatory practices was quickly expanded into a device by which employers were held liable for discrimination whenever they utilized employment criteria that produced statistical disparities. This evolution progressed from the assumption articulated by the Court in its 1977 *Teamsters* decision that "absent explanation, it is ordinarily to be expected that nondiscriminatory hiring practices will in time result in a work force more or less representative of the racial or ethnic composition of the population in the community from which employees are hired."[11] In light of the variable of individual preferences, this assumption is hopelessly flawed;[12] and given the range of possible explanations for statistical disparities—age, qualifications, interest, information, accessibility, education, and so on—mere statistics without more do not logically give rise to a significant inference of discrimination except in a broader *Griggs*-type context in which corroborating evidence of discrimination is supplied.

---

[6] *Id.*, p. 426.

[7] *Id.*, p. 436.

[8] *Id.*, p. 430.

[9] *Id.*, p. 431.

[10] See, e.g., the opinion of Judge Richard Posner concurring in part and dissenting in part in United States v. Town of Cicero, 786 F.2d 331 (7th Cir. 1986).

[11] International Brotherhood of Teamsters v. United States, 431 U.S. 324, 340 n.20 (1977).

[12] See, e.g., Thomas Sowell, *Civil Rights: Rhetoric or Reality?* (New York: William Morrow and Co., 1984), pp. 53–56.

Nonetheless, the EEOC and several courts below the Supreme Court level acted as if *Griggs* were a line of scrimmage from which to march the football steadily downfield. The cases established a three-part adverse impact analysis: (1) plaintiffs could establish a *prima facie* case of discrimination based solely on statistical disparities, (2) the employer would then have to prove the "business necessity" of its practices, and (3) the plaintiff could rebut such a defense by showing it was pretextual.[13] The EEOC guidelines go even further, requiring the employer to show that no alternative selection device is available that would produce less adverse impact.[14]

Despite the ease with which plaintiffs could force employers into court on purely statistical showings without any evidence whatsoever of intent to discriminate, the EEOC and several courts made it nearly impossible for employers to show business necessity. Departing from the *Griggs* standard of a "reasonable measure of job performance," lower courts required employers to demonstrate that the challenged job practice was "essential"[15] or justified by an "irresistible demand."[16]

In the context of employment tests, this standard required "validation" by test experts to show a precise correlation between the test and job performance, a process that often runs into hundreds of thousands of dollars and can prove completely impossible, notwithstanding the total absence of discriminatory intent.[17] As one district court judge complained in 1973, "Under this rigid standard, there is no test known or available today which meets the Equal Employment Opportunity Commission requirements for any industry."[18] Justice Harry Blackmun later warned, "I fear that a too-rigid application of the EEOC Guidelines will leave the employer little choice, save an impossibly expensive and complex validation study, but to engage in a subjective quota system of employee selection. This, of course, is far from the intent of Title VII."[19]

Indeed, such a result conflicts both with section 703(j) of Title VII, which precludes requiring employers to adopt racial preferences to eliminate statistical disparities, and section 703(h), which protects nondiscriminatory testing devices. Yet the fears expressed by Justice Blackmun were fully realized. As Michael Gold charges, "Quotas and adverse impact are practically synonymous. In theory, an employer can win an adverse impact case by proving that the challenged selection criterion is valid. In practice, this burden can almost never be carried, and the result is that employers are forced to hire and promote by quotas."[20]

---

[13] See Barbara Lindeman Schlei and Paul Grossman, *Employment Discrimination Law*, 2d ed. (Washington: Bureau of National Affairs, Inc., 1988), pp. 1324–25.

[14] 29 C.F.R. section 1607.

[15] Watkins v. Scott Paper Co., 530 F.2d 1159, 1168 (5th Cir.), *cert. denied*, 429 U.S. 861 (1976).

[16] United States v. Bethlehem Steel Corp., 446 F.2d 652, 662 (2d Cir. 1971).

[17] Michael Gold, "*Griggs*' Folly: An Essay on the Theory, Problems, and Origin of the Adverse Impact Definition of Employment Discrimination and a Recommendation for Reform," 7 *Indus. Rel. L. J.* 429, 460 (1985).

[18] United States v. Georgia Power Co., 3 Fair Empl. Prac. Cas. (BNA) 767, 780 (N.D. Ga.), *rev'd*, 474 F.2d 906 (5th Cir. 1973).

[19] Albemarle Paper Co. v. Moody, 422 U.S. 405, 449 (1975) (Blackmun, J., concurring in the judgment).

Employers have also routinely abandoned tests rather than defending them. A survey by the Equal Employment Advisory Council found that 82 percent of its corporate members had ceased the use of some or all tests for fear of litigation or due to the cost of validation.[21] The costs to our nation in terms of productivity and competitiveness—not to mention the principle of equal opportunity—are staggering.[22]

The wholesale abandonment of objective employment standards is bizarre in light of the objectives of Title VII. Logically, *objective* devices are less susceptible to discriminatory influences, yet adverse impact encourages employers to rely on *subjective* devices. Similarly, employers can avoid costly litigation by hiring proportionally, subverting equal opportunity policies in favor of racial quotas. Thus has adverse impact been transformed in Orwellian fashion from an important weapon to combat discrimination into a powerful engine of discrimination in the form of racial quotas.

Yet no assurance exists that this misapplication of adverse impact does much to solve the problems that disproportionately afflict minorities. By characterizing every racial disparity as discrimination that is curable by a quota, the adverse impact construct focuses on outcomes rather than on the need to give people the tools to pass tests and to satisfy objective standards. And as [former] EEOC Chairman Clarence Thomas charges, such an approach tacitly endorses notions of the "inherent inferiority of blacks . . . by suggesting that they should not be held to the same standards as other people."[23]

The Supreme Court acted decisively to harmonize adverse impact with the express purposes of Title VII in its decision earlier this year in *Wards Cove Packing Co. v. Atonio.*[24] In *Atonio,* the plaintiffs challenged an employer's entire range of hiring practices, relying solely on statistics showing a high percentage of nonwhite workers in certain other jobs and a high percentage of whites in other jobs. (The plaintiffs also challenged certain other practices on "disparate treatment" grounds, but these were not before the Supreme Court.) The Ninth Circuit Court of Appeals had ruled the plaintiffs' statistical showing adequate to establish a *prima facie* showing of discrimination, and required the employer to prove the business necessity of its practices.[25]

The Supreme Court reversed in a 5–4 decision written by Justice Byron White. The Court first focused on the use of statistics in establishing a *prima facie* case, and concluded that the comparison of one category of jobs with different jobs was not probative of discrimination. Rather, the Court ruled, the plaintiffs must produce statistics with respect to "'the pool of

---

[20] Gold, p. 457.

[21] Edward E. Potter, ed., *Employee Selection: Legal and Practical Alternatives to Compliance and Litigation,* 2d ed. (Washington: National Foundation for the Study of Equal Employment Policy, 1986), p. 215.

[22] See, e.g., Potter, pp. 315-19.

[23] Quoted in *Changing Course,* p. 63.

[24] 109 S.Ct. 2115 (1989).

[25] *Id.,* p. 2117.

qualified job applicants' or the '*qualified* population in the labor force'" to prepare a foundation for a showing of possible discrimination.[26] Otherwise, Justice White explained:

> any employer who had a segment of his work force that was—for some reason—racially imbalanced, could be hauled into court and forced to engage in the expensive and time-consuming task of defending the "business necessity" of the methods used to select the other members of his work force. The only practicable option for many employers will be to adopt racial quotas, insuring that no portion of his work force deviates in racial composition from the other portions thereof; this is a result that Congress expressly rejected in drafting Title VII.[27]

Moreover, the Court held, plaintiffs may not challenge the statistical "bottom line" of a range of employment practices, but must focus on the specific employment practices that produced the adverse impact. A converse rule, the Court observed, "would result in employers being potentially liable for 'the myriad of innocent causes that may lead to statistical imbalances in the composition of their work forces.'"[28] In other words, the employer in a purely statistical challenge cannot be forced to defend every single one of its employment practices, but only those that are potentially discriminatory.

The Court then turned to the employer's burden once a *prima facie* showing is made, a burden the Court characterized not as one of proof but of "producing evidence of a business justification," since as the Court noted the "burden of persuasion . . . remains with the disparate-impact plaintiff."[29] The Court emphasized that such evidence need only show that the "challenged practice serves, in a significant way, the legitimate employment goals of the employer," rather than that the practice is "'essential' or 'indispensable,'" a standard that "would be almost impossible for most employers to meet."[30] Plaintiffs would remain free to rebut such evidence by showing that alternative practices exist that would equally serve the employer's objectives, which would suggest the employer's justifications were pretextual.[31]

Atonio thus leaves intact adverse impact as a method of proving discrimination, but requires that the statistics presented actually raise a plausible inference of discrimination. Combined with the Court's recent decisions subjecting governmentally imposed racial quotas to the strictest (and almost invariably fatal) constitutional scrutiny,[32] *Atonio* makes clear that the Court will no longer accept racial quotas as a superficial substitute for equal employment opportunity.

---

[26] *Id.*, p. 2122 (citation omitted).

[27] *Id.*

[28] *Id.*, p. 2125 (citation omitted).

[29] *Id.*, p. 2126.

[30] *Id.*, p. 2125-2126 (citation omitted).

[31] *Id.*, p. 2126.

[32] See City of Richmond v. J.A. Croson Co., 109 S.Ct. 706 (1989) and Wygant v. Jackson Bd. of Educ., 476 U.S. 267 (1986).

The Mischief Continues

Despite the Supreme Court's rulings, efforts to use the coercive apparatus of the state to advance the antistandards and pro-racial quota agenda—regardless of the perverse consequences that may result—continue unabated. Two examples will illustrate these efforts.

The first is an invidious and profoundly unlawful practice engaged in by the United States Employment Service (USES) called "race norming." The USES coordinates the job referral programs of State employment services nationwide. It uses as a screening device the General Aptitude Test Battery (GATB), which the Labor Department has defended as valid. Nonetheless, since the GATB produces some adverse racial impact, the USES has constructed a "within-group score conversion" process that adds points to the scores of applicants from certain specified groups in order to assure proportional job referrals. In other words, after adopting a test battery it considers the best at predicting job performance and hence assuring merit-based job referrals, the USES deliberately distorts that process with a racial quota system.

Earlier this year, a panel of the National Academy of Sciences attempted to place a scientific veneer on the practice of making score adjustments on the basis of race. In its published study,[33] the panel found that GATB is a good predictor of job performance; that it, therefore, has a positive effect on productivity; and that it is not racially biased and may in fact *overpredict* performance for blacks. Such questions were the extent of the panel's mandate. Nonetheless, in a remarkable display of social engineering over science, the panel concluded that since certain groups attain higher scores on GATB than others, score adjustments are appropriate, a conclusion that has been severely criticized.[34]

The USES's race-norming policy is clearly unlawful. Unless the test battery is discriminatory—and even the National Academy of Sciences panel concluded it was not—no justification exists to adulterate it in a manner that apportions opportunities on the basis of race or gender. Even if the test was biased, the proper remedy would be to fix the problem or develop a better test rather than to superimpose a permanent racial quota system like race-norming. I am very disappointed that a group of purported scientific experts would place its imprimatur on such an obviously flawed, quick-fix, nonsolution.

A second illustration of the departure from the principles embodied in the Constitution and our civil rights laws is California's blacks-only ban on I.Q. tests, the policy we are challenging in *Crawford v. Honig*. This policy was adopted in response to an earlier lawsuit challenging as discriminatory against blacks the use of I.Q. tests by public school systems to assign students to special education classes. We take no position on that earlier lawsuit, nor on the State's decision to proscribe the use of I.Q. tests for that purpose. To be sure, the State must exercise extraordinary care to use the best devices available so as to ensure that only those children who belong in special education classes are assigned there, and certainly that racial considerations play absolutely no part in that process.

---

[33] John A. Hartigan and Alexandra K. Wigdor, eds., *Fairness in Employment Testing: Validity Generalization, Minority Issues, and the General Aptitude Test Battery* (Washington: National Academy Press, 1989).

[34] See, e.g., Jan H. Blits and Linda S. Gottfredson, "Equality at Last, or Lasting Inequality? Race-Norming in Employment Testing," *Society* (in publication); "More Normal Nonsense," *Fortune* (July 17, 1989), p. 118.

But California's policy went much further than that. While public school districts remained free to provide I.Q. tests for other diagnostic purposes, blacks were prohibited from taking them. Thus, when Mrs. Mary Amaya attempted to arrange with her local school district an I.Q. test for her son Demond Crawford in order to determine that intelligence was *not* the source of his school problems, she was told she could no do so because Demond's skin color is black. Since Demond is half Hispanic, however, Mrs. Amaya was advised that he could take the test if she would reclassify him as Hispanic. Such a suggestion conjures images of Adolph Plessy, who was forced to ride in the "colored" section of a railway car during the Jim Crow era because he was 1/12 black.[35]

That we continue to assign opportunities solely on the basis of race—that we continue to deprive people from making informed judgments on their own behalf based on patronizing and paternalistic assumptions—is testimony to how far we have strayed from the principle of nondiscrimination that animated our civil rights laws. We cannot deliver on our nation's promise of equal opportunity until we purge such notions from our system once and for all.

Missed Opportunities?

Racial quotas and the abandonment of tests and other standards are surface-deep remedies that distract us from the important task of securing for all Americans truly equal opportunities. What we ought to be doing is trying to find ways to help disadvantaged individuals pass tests and satisfy objective standards.

In this era of serious shortages of skilled labor, the time is ripe for approaches that focus on human capital development and economic mobility. Between now and the year 2000, two out of every three new work force entrants will be female or minority. Opportunities abound like never before for individuals outside the economic mainstream to earn their share of the American Dream. But many such individuals—for reasons ranging from inadequate jobs skills to poverty to discrimination to inferior schooling to ghetto isolation—lack the ability to take advantage of those opportunities. Affirmative action designed to bridge these gaps will make a far bigger difference than quotas ever have in expanding meaningful employment opportunities for the most truly disadvantaged in our society.

I have profiled a number of such approaches—what I call "proactive" affirmative action—in a recent study for the Department of Labor entitled *Opportunity 2000: Creative Affirmative Action Strategies for a Changing Work Force.*[36] In this study, my coauthor and I explore ways of bringing into the work force in a productive way members of groups that have not been fully included in the past: minorities, economically disadvantaged, women, older workers, and the handicapped. None of the approaches involve quotas or the abandonment of standards. Rather, they focus on investing in human capital development and in expanding economic mobility. Most importantly, unlike quotas, they expand the pie rather than merely redistribute it.

---

[35] See Plessy v. Ferguson, 163 U.S. 537 (1896).

[36] Washington, DC: U.S. Department of Labor, 1988.

This does not mean that we should in any way shortchange the effort to eradicate barriers—tests included—that are arbitrary or discriminatory. Indeed, an entire array of arbitrary government-imposed barriers to economic, educational, and entrepreneurial opportunities exists that we have not yet begun effectively to attack.[37] We ought to focus considerably more attention to eradicating obstacles that prevent individuals from controlling their own destinies, such as excessive regulations on entry-level economic activities, the public school monopoly, the welfare system, and crime.

To summarize, the assault on testing is not the same as an assault on discrimination; indeed, it often operates at cross-purposes with such an effort. Eradicating all tests will not aid the cause of equal opportunity or of minority advancement. Rather, it will make us a less productive society, one that applies subjective criteria (such as race) instead of objective measures in apportioning opportunities. That is precisely the opposite result intended by the civil rights movement that produced *Brown v. Board of Education* and the Civil Rights Act of 1964. Let's not turn our backs on the dream when we are on the threshold of making it a reality.

---

[37] See Clint Bolick, *Changing Course: Civil Rights at the Crossroads* (New Brunswick, NJ: Transaction Books, 1988); Walter Williams, *The State Against Blacks* (New York: McGraw-Hill Book Co., 1982).

Tests are "Useful Servants," Not the "Masters of Reality"

By Barry L. Goldstein*

The debate over the use of tests in employment is often characterized by hyperbole. For example, in 1984 the then-Chairman of the EEOC, Clarence Thomas, stated that *Griggs* "has been overextended and overapplied." He continued by pointing out that "y[ou] get people now saying if you don't have a certain number of women or blacks on the job then you are guilty of discriminating. [For example,] if it's an engineering job and [you] have a certain number of blacks because few blacks have engineering degrees, there are people who want to ask if you . . . need an engineering degree. . . . That's going too far."[1] Mr. Thomas, who had considerable positive accomplishments during his tenure at the EEOC, really fell down on his simplistic criticism of testing law. The Chairman created a straw-person argument that has nothing to do with reality. In the many volumes of fair employment decisions there is not a single decision that seriously questions the use of engineering qualifications for an engineering job.

There is an extremely serious social reality underlying the debate on the use of tests in making employment decisions. We should not lose sight of that social reality and indulge in simplistic notions about the use and worth of tests nor about the proper reach and effect of fair employment law. As stated in the report on testing issued in 1982 by the National Research Council and the National Academy of sciences:

> The salient social fact today about the use of ability tests is that blacks, Hispanics, and native Americans do not, as groups, score as well as do white applicants as a group. When candidates are ranked according to test score and when test results are a determinant in the employment decision, *a comparatively large fraction of blacks and Hispanics are screened out*. . . .
>
> So long as the[se] groups . . . *continue to have a relatively high proportion of less education and more disadvantaged* numbers than the general population, those social facts are likely to be reflected in test scores. That is, even highly valid tests will have adverse impact.[2]

There are academics and some testing professionals who look at these test score differences and state, in effect, that these scores reflect serious group differences. For example, Professor Linda S. Gottfredson states that "current black-white differences in test scores must be taken seriously [because [t]hey represent real differences in the capacity to learn and perform well a wide variety of job tasks in a wide range of jobs; [these differences are] stubborn and so are likely to be with us for some time to come; *and their impact on job success is not effectively short-*

---

* In June 1989, when I made the oral presentation to the Commission, I was director of the Washington Office of the NAACP Legal Defense & Educational Fund, Inc. At present, I am a partner with the law firm of Saperstein, Mayeda, Larkin & Goldstein in Oakland, California.

[1] "EEOC Chief Cites Abuse of Racial Bias Criteria," *Washington Post* (Dec. 4, 1984) at A-13.

[2] Committee on Ability Testing, National Academy of Science/National Research Council, *Ability Testing: Uses, Consequences and Controversies*, 143, 146 (1982).

circuited by education, training, or experience."[3] Failure to follow test scores, we are told by Gottfredson and others,[4] will result in the loss of untold billions of dollars in productivity and will endanger America's competitive position in the world economy.

These proponents of the widespread use of testing unfettered by the need to justify that use by the demonstration of a business necessity as required under the adverse impact standard are mistaken. If their advice is followed, significant benefits gained from the implementation of the fair employment law will be lost. Similarly, if the Congress does not restore the legal standards that were in effect prior to the Supreme Court's decision in *Wards Cove Packing Co. v. Atonio,*[5] equal employment opportunity in the workplace will be seriously harmed.

When Congress passed Title VII of the Civil Rights Act of 1964, it reversed the failures in our country's commitment to fair employment opportunity that occurred after the Civil War and World War II. In *Wards Cove*, the Supreme Court has sounded the call to a third retreat from effective civil rights enforcement. The call to retreat must be rejected.

Title VII has contributed to the expanding job opportunities for minorities and the removal of discriminatory barriers. "Nearly a quarter of the minority labor force of 1980 were in significantly better occupations than they would have been under the occupational distribution of 1965."[6] In a comprehensive analysis of the effect of Title VII, Professor Jonathan Leonard determined that the implementation of the antidiscrimination law from 1966 to 1977 significantly raised the share of employment opportunities, pay, and job levels of black workers without any "significant effect on productivity."[7]

---

[3] Gottfredson, "Reconsidering Fairness: A Matter of Social and Ethical Priorities," 33 *Journal of Vocation and Behavior*, 293, 299-30 (1988) (emphasis added).

[4] *See also*, Schmidt, "The Problem of Group Differences in Ability Test Scores in Employment Selection," 33 *Journal of Vocational Behavior*, 272 (1988); and Scharf, "Litigating Personnel Measurement Policy," 33 Journal of *Vocational Behavior*, 235 (1988).

[5] 109 S.Ct. 2115 (1989).

[6] Blumrosen, "The Group Interest Concept, Employment Discrimination, Legislative Intent: The Fallacy of *Connecticut v. Teal,*" *20 Harvard Journal on Legislation* 99 (1983).

[7] Leonard, "Anti-discrimination or Reverse Discrimination: The Impact of Changing Demographics, Title VII, and Affirmative Action on Productivity," 19 *The Journal of Human Resources* 145 (1984).

Professor Richard Freeman has described Leonard's study as the "only significant empirical study" of the effect of fair employment laws on productivity. Freeman, "Affirmative Action: Good, Bad or Irrelevant?" *New Perspectives* (1984: U.S. Comm. on Civil Rights).

In testimony before the U.S. Commission on Civil Rights, Professor Leonard further described the findings of his study:

> Relative minority and female productivity increased between 1966 and 1977, a period coinciding with government anti-discrimination policy to increase employment opportunities for members of these groups. *There is no significant evidence here to support the contention that this increase in employment equity has had marked efficiency costs.* The relative marginal productivities of minorities and women have increased as they have progressed into the work force, suggesting that discriminatory employment practices have been reduced.

Leonard, Testimony before the U.S. Commission on Civil Rights, (February 1985) (emphasis added).

BEST COPY AVAILABLE

In 1971 the Supreme Court decided "the most important court decision in employment discrimination law,"[8] *Griggs v. Duke Power Co.*[9] The Commission is familiar with *Griggs* and I will only review briefly the decision. The Court determined that "Congress directed the thrust of the Act to the *consequences* of employment practices, not simply the motivation."[10]

Critically, the Court interpreted Title VII in a *practical* manner. If a plaintiff demonstrates that a device or system "selects applicants for hire or promotion in a racial pattern significantly different from that of the pool of applicants," then the selection system is illegally discriminatory unless the employer meets the burden of showing that any given requirement is necessary. "The touchstone [for this determination] is business necessity."[11]

This is a uniquely practical approach to removing discriminatory barriers. The focus remains upon the selection system; the case does not depend upon the "intent" or "state of mind" of the employer. Severe barriers to equal employment opportunity may not remain unassailable because the plaintiff is unable to show that the employer acted in "bad faith." Moreover, the justification for the continuation of the barriers falls upon the employer, the party who has access to the requisite evidence and who, as a matter of course, or good business practices, should have a justification for the use of a selection system. The Court stated as follows:

> The facts of this case demonstrate the inadequacy of broad and general testing devices as well as the infirmity of using diplomas or degrees as fixed measures of capability. History is filled with examples of men and women who rendered highly effective performance without the conventional badges of accomplishments in terms of certificates, diplomas, or degrees. Diplomas and tests are *useful servants*, but Congress has mandated the commonsense proposition that they are *not to become the masters of reality*.[12]

This practical and fair approach taken in this unanimous Supreme Court opinion received widespread support.

Almost immediately after the Supreme Court issued the *Griggs* opinion, Congress recognized the importance of the opinion, well described that importance, and determined that the *Griggs* principles should be extended to Federal Government employment.[13]

> [The Civil Service Commission] apparently has not fully recognized that the *general rules and procedures that it has promulgated may in themselves constitute systematic barriers to minorities and women.* Civil Service selection and promotion techniques and requirements are replete with artificial requirements that place a premium on "paper" credentials. Similar requirements in the private sectors of business have often proven of questionable value in predicting job performance and have often resulted

---

[8] B. Schlei and P. Grossman, *Employment Discrimination Law* (1983) at 6.

[9] 401 U.S. 424 (1971).

[10] *Griggs*, 401 U.S. at 431.

[11] *Id.*

[12] *Griggs v. Duke Power* Co., 401 U.S. at 433 (emphasis added.)

[13] As originally enacted, Title VII only applied to private employment. The Equal Employment Opportunity Act of 1972 extended the law to Federal, State, and local government employment.

> in perpetuating existing patterns of discrimination (See, e.g., *Griggs v. Duke Power* . . . .) The inevitable consequence of this kind of a technique in Federal employment, as it has been in the private sector, is that classes of persons who are socioeconomically or educationally disadvantaged suffer a heavy burden in trying to meet such artificial qualifications.
>
> *It is in these and other areas where discrimination is institutional, rather than merely a matter of bad faith, that corrective measures appear to be urgently required. For example, the Committee expects the Civil Service Commission to undertake a thorough re-examination of its entire testing and qualification program to ensure that the standards enunciated in Griggs are fully met.*[14]

As the Senate Committee perceptively described in 1971, the "full" implementation of the *Griggs* principles is critical to meeting the fundamental goal of equal employment opportunity. Nineteen years later the U.S. Commission on Civil Rights should stay the course that was well-chartered by *Griggs* and the 1972 Congress and support the Civil Rights Act of 1990 in order to overturn the limitations on the *Griggs* principles placed by the Supreme Court in *Wards Cove Packing Co.*

Before turning to the specific problems created by *Wards Cove Packing Co.*, it is useful to turn to several specific examples of the types of job opportunities in the 1970s that were opened to blacks on a fairer basis after the *Griggs* decision. Two types of jobs provide adequate illustration: police officer and craft worker. For both types of jobs, selection devices, such as tests or referral practices, had served to limit opportunities of blacks. In the 1970s, after the *Griggs* opinion, the number of blacks working in these job categories increased dramatically.

In 1972 blacks made up 3.2 percent or 15,872 of the 496,000 electricians in the country, whereas in 1979 blacks represented 5.6 percent or 35,490 of the 640,000 electricians in the country.[15]

In general, during the period 1972 through 1979, the number of blacks employed in the craft and kindred census category, increased by 270,000.[16]

In 1970, 6.4 percent or 23,796 of the 375,494 police officers and detectives in the country were black,[17] whereas in 1982, 9.3 percent or approximately 47,000 of the 505,009 police officers in the country were black.[18]

While several factors contributed to the substantial increases during the 1970s in the number of blacks working in craft, police, and similar positions, the effective implementation of Title VII and application of the *Griggs* rules contributed substantially.

---

[14] S. Rep. No. 92-415, 92nd Cong., 1st Sess. at 14-15 [emphasis added.]

[15] 1980 *Statistical Abstract of the United States* (1980) at table 697.

[16] *Id.*

[17] U.S. Bureau of the Census, *Census of the Population: 1970. Vol. 1, Characteristics of the Population, Part 1, United States Summary—Section 1* (1973) at table 223.

[18] *1984 Statistical Abstract of the United States* (1984) at table 696.

What did the Supreme Court do in *Wards Cove* that so undermined the *Griggs*[19] principle that Congress should act promptly to reverse? In short, in my view as a litigator who has sought to enforce fair employment law for almost 20 years, the Supreme Court has made the *Griggs* impact standard largely ineffective. Private attorneys, who litigate the overwhelming majority of fair employment cases, would find it difficult, if not impossible, to litigate fair employment cases under the *Griggs* impact standard as changed by the Supreme Court in *Wards Cove Packing Co.* Private attorneys, like myself, will continue to take and litigate cases of intentional discrimination. However, the "most important" fair employment decision, *Griggs v. Duke Power Co.*, *supra*, which has served to open job opportunities to disadvantaged minorities, is reduced to a minor role in the enforcement of fair employment law by the Supreme Court in *Wards Cove Packing Co.*

Let me discuss three aspects of the *Wards Cove Packing Co.* opinion that thwart the use of the *Griggs* impact standard to remove unnecessary barriers to fair employment: (1) the proof that a practice disproportionately limits the opportunities of minorities or women, the plaintiffs' *prima facie case*; (2) the burden of proof; and (3) the standard for justifying a selection practice or system that has an adverse impact.

Prima Facie Case

There are two principal aspects to the *prima facie* or adverse impact analysis in *Wards Cove Packing Co.* The first aspect, the proper measure of the relevant labor pool, is not objectionable. This aspect of the *Wards Cove* decision remains unchanged by the proposed legislation. The second aspect, the so-called "pinpointing" requirement, is objectionable.

If an analysis of the actual applicant flow is impossible or inappropriate, then the plaintiff may seek to *demonstrate* adverse impact by reference to the *relevant qualified labor pool.*

The plaintiff must demonstrate that the labor pool reflects *qualified* and *available* workers. As a general matter, the Supreme Court did not alter existing law with respect to the labor force analysis. Thus, in the majority opinion, Justice White correctly stated that it is "nonsensical" to compare the proportion of minorities in the general work force with the proportion of minorities selected for skilled positions, such as "boat captains, electricians, doctors, and engineers."[20] Similarly, in the dissenting opinion, Justice Stevens correctly concluded that "[a]n undisputed requirement for employment either as a cannery or noncannery worker is availability for season employment in the far reaches of Alaska."[21] Accordingly, any analysis of the relevant labor pool need include a reasonable analysis of workers available for seasonal work.

The requirement for an appropriately relevant labor market analysis is a two-edged sword. Qualification standards may *increase or decrease* the proportion of minorities in the relevant labor pool. For example, the proportion of black doctors or engineers is smaller than the proportion of blacks in the labor force; thus, the relevant *qualified* black labor force for medical or engineering positions would be *reduced* by a qualification requirement. On the other hand, if professionals and highly skilled workers are removed in determining the available labor pool for unskilled positions, the relevant *qualified* black labor force would be *increased* by properly adjusting for relevant requisite qualifications. Similarly, since proportionally more minorities are available than whites for seasonal work, the proper adjustment of the relevant labor pool, as suggested by Justice Stevens, would *increase* the relevant qualified minority labor force by adding a qualification requirement.

---

[19] I have referred to the *Griggs* principles; however, these principles have regularly been applied in other Supreme Court opinions, see, e.g., *Albermarle Paper Co.*; *Dothard v. Rawlinson*, 433 U.S. 321 (1977); *Connecticut v. Teal*, 457 U.S. 440, 446 (1982), and in hundreds of lower court cases.

[20] *Wards Cove Packing Co.*, 109 S.Ct. at 2122.

[21] 109 S.Ct. at 2134.

If a plaintiff challenges a requirement, such as a medical, engineering, or other undisputedly relevant degree, electrical or boat pilot license, just to name several qualifications that were apparently appropriate in *Wards Cove Packing Co.*, then either the applicant flow or the labor pool must be adjusted for these qualification requirements when the adverse impact analysis is made. The *Wards Cove* requirement with respect to a proper labor force analysis is not objectionable. Thus, the focus properly turns to whether the selection practices, which are in dispute, have disproportionately excluded minorities from job opportunities. In *Wards Cove* the practices to which the adverse impact analysis should have applied included nepotistic hiring, word-of-mouth recruiting, and subjective decisionmaking.

These types of practices, which may often serve as unnecessary or even deliberate barriers to the job opportunities of minorities and women, should be subject to challenge under the *Griggs* adverse impact analysis when the plaintiffs prove by a preponderance of the evidence that these practices *combine* to limit opportunities of minorities or women. *Wards Cove Packing Co.* wrongly insulates these practices. The Court ruled that even if the plaintiffs *properly* showed by reference to the *relevant qualified* labor pool that the selection practices had an adverse impact, "this alone will *not* suffice to make out a prima facie case of disparate impact."[22] Thus, even if plaintiffs conclusively demonstrated that 40 percent of the qualified labor pool were minorities and that only 10 percent of the persons selected after the operation of three selection practices, nepotism, word-of-mouth recruiting, and subjective evaluation process, the plaintiffs would still fail to show adverse impact. The plaintiffs must identify and prove which one of the three practices caused the impact.[23]

This burden remains on the plaintiffs even though the employer has the best access to the relevant evidence, has a duty under appropriate regulations to keep the relevant data, and, most importantly, even though there is *no dispute* that the employer's selection system serves as a possibly illegal barrier to equal job opportunity. This "pinpointing" requirement is an improper impediment to the enforcement of the fair employment law; it is comparable to sending players off on a treasure hunt without any clues.

Prior to filing a lawsuit, a plaintiff and his or her attorney may have substantial evidence that an overall selection practice has adverse impact. It is possible to ascertain by observation some sense of the proportion of minorities in the applicant pool and the proportion of minorities selected for the work force; or alternatively, with the assistance of a labor market economist, it is possible to make a good estimation of the proportion of minorities in the relevant qualified labor force.

---

[22] 109 S.Ct. at 2125.

[23] *Id.*

However, it is not possible to know prior to filing suit whether the employer has maintained adequate records in order to identify which particular selection practice or practices have caused the adverse impact. Thus, even where there is substantial evidence of adverse impact caused by the selection system and even where the system contains practices, such as nepotism, word-of-mouth, and subjective decisionmaking, that frequently have been used to discriminate illegally, the plaintiff or plaintiff's lawyer may conclude that the result of a lawsuit under the *Wards Cove* pinpointing standard is too uncertain to litigate. They may understandably decide not to embark upon a treasure hunt without clues. The effective implementation of Title VII is harmed by the pinpointing requirement of *Wards Cove*. The Civil Rights Act of 1990 properly removes this requirement.

Burden of Proof

It cannot be seriously disputed that prior to *Wards Cove Packing Co.* the burden of persuasion was placed squarely on the employer to show that its use of a selection practice that disproportionately limited the opportunities of minorities or women was justifiable. As the Court simply stated in *Griggs*: "Congress has placed on the employer the *burden of showing* that any given requirement must have a manifest relationship to the employment in question."[24] From 1972 through 1988 I litigated employment cases in many parts of the United States; in not a *single* instance did a defendant or Court suggest that the burden of persuasion did not shift under the *Griggs* rule.

The Supreme Court's ruling that "[t]he burden of persuasion . . . remains with the disparate-impact plaintiff" is a clear signal that courts are to treat fair employment plaintiffs less sympathetically and that close cases should be decided against claimants. Moreover, it is more difficult for three practical reasons for plaintiffs to prove that a practice is *not* justifiable, than for a defendant to prove a practice justifiable.

First, the defendant has access to the information about the job and selection practice in question. After all, the employer chose the practice in the first place; the employer knows the reason for its decision. Second, it is easier to prove the affirmative, that a practice is justified by business necessity, than to prove the negative. This is especially true given the lax standard under *Wards Cove Packing Co.* for making this showing. Third, the employer has more experience and resources to show that a selection practice is required by business necessity than a plaintiff has to show the negative.

By reversing the long-established *Griggs* burden-shifting rule, the Supreme Court in *Wards Cove Packing Co.* sent a clear message—it will be difficult, if not impossible, to win many legitimate employment discrimination claims. The message will be heard; unless Congress reverses this result, legitimate claims will not be pursued or, if pursued, many legitimate claims will be lost.

---

[24] 401 U.S. at 432 (emphasis added).

Standard of Proof

For 18 years the courts, employers, litigants, and governmental agencies have followed the classic standards set forth in *Griggs:* an employer must demonstrate that the employment practice shown to have resulted in disparate impact was justified by a business necessity. The *Griggs* Court's made clear with strong language—"business necessity"—that fair employment opportunity was important and that barriers to the hiring or advancement of minorities and women would be closely scrutinized.

In *Wards Cove Packing Co.* the Supreme Court almost parodies these standards. The Court changes the "touchstone. It is no longer "business necessity;" rather "[t]he touchstone of this inquiry is a reasoned review of the employer's justification for his use of the challenged practice."[25] What does this mean? The Supreme Court states that the Court requires more than "[a] mere insubstantial justification,"[26] but less than "essential" or "indispensable."[27] This is an enormous playing field without much guidance or many rules provided.

Why did the Court jettison the 18 years of interpretation of the *Griggs* standard? As described earlier, the *Griggs* rules have demonstrably worked to increase fair employment opportunity. Moreover, as shown in Professor Leonard's study, there is no evidence that the gains made by minorities harmed productivity. In fact, the executive officer of the American Psychological Association, Dr. Goodstein, stated in congressional testimony "that psychologists generally agree that the caliber of employment practices in organizations *has improved dramatically* since publication of the existing Uniform Guidelines[28] in 1978."[29]

Even more important, the courts along with the Federal agencies have fleshed out the *Griggs* principles over an 18-year period. The predictability and guidance achieved by the administrative agency and court decisions are lost by the dramatic change in the standard made by the Supreme Court in *Wards Cove Packing Co.*

It is critical to restore the case law and predictability which was overturned by *Wards Cove Packing Co.* It is difficult for attorneys to undertake the representation of potential victims of discrimination when there is unpredictability in the law.

---

[25] 109 S.Ct. at 2126.

[26] It is incredible that the Court even has to say that "a *mere* insubstantial justification" is inadequate. Could "a mere insubstantial justification" ever be adequate for anything?

[27] *Id.*

[28] The Uniform Guidelines on Employee Selection Procedures, 29 C.F.R. §1607, were promulgated, by the Federal agencies, Equal Employment Opportunity Commission, Departments of Justice and Labor, and the Civil Service Commission (now the Office of Personnel Management), charged with enforcing the fair employment laws. The Guidelines were drafted in order to establish specific standards for implementing the *Griggs* adverse impact principle.

[29] Goodstein, "On the Subject of Uniform Guidelines on Employment Selection Procedure," Testimony before the Subcommittee on Employment Opportunities of the Education and Labor Committee of the House of Representatives (Oct. 2, 1985) (emphasis and footnote added).

For example, it is my view that under the *Wards Cove* rules a case with the same facts as *Griggs* might likely be decided for the defendant. The Duke Power Company had not intended to discriminate; in fact, the company had engaged in special efforts to assist undereducated employees, black and white.[30] Moreover, the high school education requirement, which was struck down by the Supreme Court, was used by the company for the selection of employees into departments with skilled jobs, such as machinist, electrician, welder, power station operator, and lab technician.[31] The Duke Power Company *never* required a high school diploma for the Labor Department where jobs requiring manual work were located.[32] It is certainly arguable that under the *Wards Cove* standard, where the "touchstone" is a "reasoned inquiry" rather than "business necessity" and where the plaintiff rather than the defendant has the burden of proof, that a court might determine that a high school diploma was a "legitimate" requirement for these skilled jobs in a power plant. Therefore, if *Wards Cove* principles had applied in 1971, Willie Griggs and other black workers would never have had the opportunity to work in jobs commensurate with their actual abilities and skills.

Conclusion

The U.S. Commission on Civil Rights should urge Congress to overturn the *Wards Cove* rules in order to continue this country's commitment to fair employment opportunity. We should not retreat from that commitment as we did after the Civil War and after World War II. Before we can address additional difficult questions regarding the use of employment tests, the *pre-Wards Cove* standards defining fair employment law must be reestablished.

But there is a further practical national interest compelling the restoration of the *Griggs* rule. "[B]etween now and the year 2000" non-whites "will make up 29 percent of the new entrants into the labor force . . . twice their current share of the work force."[33] "Almost two-thirds of the new entrants into the work force between now and the year 2000 will be women."[34] Given the fact that, as a whole, the work force during this period "will grow more slowly than at any time since the 1930s," there is a compelling need to "integrate (female,] Black, and Hispanic Workers fully into the economy."[35] Part of this integration must occur through training and education; but another part should be the removal of unnecessary barriers to the employment opportunity of minorities and women. The *Griggs* standards aim towards this goal; *Wards Cove Packing Co.* is a detour. Congress should return the country to the path well-marked by Congress in 1964 and 1972 and by the Supreme Court in *Griggs.*

---

[30] 401 U.S. at 428–29.

[31] The jobs affected by the education and testing requirements are described in the district court opinion. *Griggs v. Duke Power Co.*, 292 F. Supp. 243, 245 n.1 (M.D. N. C. 1968).

[32] 401 U.S. at 427.

[33] *Workforce 2000 Work and Workers for the 21st Century* (1987) at xx. This report was prepared by the Hudson Institute for the Department of Labor.

[34] *Id.*

[35] *Id.* at xiv.

Analysis

The positions of each of the panelists expressed in their papers and during the consultation are summarized below. Then, areas of agreement and disagreement are summarized. Finally, the major findings are listed.

Dr. D. Monty Neill, Associate Director, National Center for Fair & Open Testing (FairTest)

Neill espouses the FairTest goals of enhancing equity and enabling access. He believes that tests, as currently constructed and used, create unfair barriers to achieving these goals. He points out that testing in public schools has increased, especially in school districts where low-income and minority students are concentrated. In education, tests sort students into classrooms with inequities in educational services; they narrow school curricula and force schools to overemphasize basic skills rather than critical thinking, reasoning, and problem solving; they shift control and authority from teachers, parents, and the community to the testing industry; and they discourage students, causing them to drop out. In employment, tests exclude qualified applicants, particularly minorities, hurting both the applicants and the industries. These harmful social effects are sufficient, Neill argues, to reject the use of standardized, multiple-choice tests for most purposes.

Neill believes most tests are not fair or objective. First, they represent mental development as a single dimension or number, rather than as multiple facets of knowledge, learning, and thinking. Second, the unreliability of tests can produce score differences spanning cut points with dramatically different impacts on test takers (e.g., college admission or its denial). Finally, even good tests do not predict later performance very well, certainly not well enough to warrant making decisions solely, or even primarily, by test scores. Poor predictions occur, particularly when expectations for performance differ from those for which the test was developed or when the selection procedure creates a self-fulfilling prophecy.

Furthermore, he believes validation procedures are inadequate. Content validity studies must address both what content should and should not be included. Test items should be balanced to cover adequately the range of content of what is taught or done on the job. They should reflect the diversity of language, experience, and perspectives of the test takers. They should also be subjected to disconfirming hypotheses, for example, do those who fail the item lack adequate content knowledge to be good teachers?

Relying upon comparisons to other tests, he suggests, is not sufficient to demonstrate criterion-related or construct-related validity because the comparison depends upon the validity of the other test. Furthermore, the validity of tests should be compared to the validity of teacher judgments or other high-quality alternatives and not to random chance, as is often done.

Tests should measure what they claim to measure. Construct validity studies must also examine the relationships among theories of knowledge, ability, and performance; tests; and test use. The concept of construct validity must be expanded to consider social or educational values and the effects of test use.

Bias creeps into the questions themselves through the language of the tests, through differences in cultural experience and perspective or in ways of knowing and problem solving, and through the timed format. Procedures to identify biased items often eliminate items upon which minorities do better because the items behave oddly in the majority-white sample. Test developers, however, do not routinely eliminate items that contain bias.

Direct evidence of traits, referred to as "authentic" assessments, have been and are being developed and can be used instead of tests. Work samples and portfolios are some examples of authentic assessments.

Finally, FairTest would require that test developers give educators, schools, test takers, and independent researchers access to previous test questions, their answers, and the descriptive and statistical data on test construction and validation so they may verify test publishers' claims. The Federal or State government should establish guidelines for the testing industry, require information on standardized tests to be made public, analyze test results to guard against bias, and set standards for the proper use of test results.

Dr. James W. Loewen, Professor of Sociology, University of Vermont

Loewen argues that differences in test scores emanate from the social structure. Some differences in social structure (e.g., differences in the race of test administrators and test takers, in access to coaching, and in familiarity with words) produce a bias in test results that should be eliminated. Other differences in social structure (e.g., unequal school finance, differences in prenatal care and nutrition, and differences in expectations and attitudes perpetrated by occupational segregation) affect test scores legitimately, but should not be allowed to legitimize group differences in scores. He is concerned that an emphasis on test scores directs attention to individualistic solutions rather than to changes in the social structure.

Thus, Loewen accepts that aptitude tests show adverse impact, some of which is not bias. But access to college education should not depend upon test scores that themselves largely depend upon race, or income, gender, and place of residence, he believes. Furthermore, the Nation can eliminate inequity and increase opportunities with policy changes that capture this approach.

Because the social structure creates unequal opportunities, Loewen believes affirmative action is necessary. By affirmative action he means admitting a cross section of America, but perhaps chosen by meritocratic means within each group. Tests should be designed and validated accordingly. Specifically, he suggests (1) items with the most adverse impact should be dropped (i.e., "the Golden Rule procedure"), even though the differences they reveal may be valid; (2) studies of predictive validity should be conducted on individual test items within gender and racial groups; and (3) average test scores should be balanced by adding a constant for members of low-scoring groups, with both adjusted and unadjusted scores reported to test takers and users.

Through strategies such as these, he argues, test bias is the easiest source of adverse impact to remedy (i.e., easier than, say, major prenatal care programs or massive changes in taxation methods). Yet, the Educational Testing Service uses none of them on the Scholastic Aptitude Test. The gender gap on the verbal SAT can be eliminated; the gender gap on the math SAT can be reduced by about one-third. The black-white gap on the verbal SAT can be cut by about 40 percent and the math gap by perhaps a third by applying the Golden Rule procedure.

Loewen opposes the Educational Testing Service's use of Differential Item Functioning (DIF) and other similar methods of identifying biased items.[1] Methods that require new test items to correlate with all the old test items or that take overall test score into account when looking at how groups perform differently build inertia into the test construction process, making change to less biased items difficult. Another concern is that researchers using these analyses will remove items that favor, as well as those that hurt, the lower scoring group.

Other comments of his agree with the FairTest position. He suggests that aptitude or ability tests are really measuring background, not what they are supposed to measure. Like Dr. Neill, he believes predictive validity—that is, the relationship between test scores and performance—is low in tests such as the Scholastic Aptitude Test. Alone, it is not adequate to claim that a test is unbiased. Finally, he too suggests that Federal oversight of test makers is necessary.

---

[1] He concurs with Nancy Cole's statement that the DIF analysis is not an analysis of bias. He suggests that the statement means DIF statistics are insensitive in identifying biased items; she, however, would suggest that DIF analysis is oversensitive and identifies unbiased items along with biased ones.

Dr. Nancy S. Cole, Executive Vice President, Educational Testing Service

Cole notes that test validation, fairness, and bias refer not just to the test and its development but to the test's use as well. Validation, she says, is the process of accumulating evidence that the inferences made from test scores are sound. Thus, validity is a characteristic of the inferences based on the test scores. Some inferences based on a test score may be sound whereas others based on the same test may not be sound. A test cannot be valid in general or for all uses.

Fairness and bias refer to a special type of validity or invalidity. Bias is invalidity with respect to certain groups. The key question about bias and fairness is whether the inferences from test scores are inappropriate or appropriate for members of a group of concern.

Validation (and the study of fairness) requires many kinds of evidence including the context of test use (e.g., whether it is used for self-evaluation, selection, or to provide an intervention); the content and format of questions and their fairness for particular groups; administration and scoring; the internal test structure (i.e., the relationship between its various parts); and the external test relationships (e.g., the relationship between SAT scores and college performance). Discussions of bias have focused mostly on differential performance on individual test items (the internal test structure) and differential predictions for different groups (the external test relationships). But validity and fairness cannot be represented by a single number from a single approach. Nor should a test score be used singlehandedly for important decisions when other information is clearly relevant.

Cole believes that group differences in test scores or test items are not necessarily a sign of bias. The scores may reflect valid differences in relevant skills or knowledge created by differences in education and opportunities. Groups that differ in education and opportunity are likely to differ on various educational accomplishments and therefore on educational test scores. Tests are not the cause of such differences in educational attainment but the result.

Group differences on a test or question appropriately trigger concern about possible bias. However, in order to infer that the test or question is biased, differences in educational attainment must be ruled out as a reason for the score difference. Thus, all attempts to measure bias in the technical literature involve some type of matching of examinees from different groups in educational attainment. If, after matching, group differences on the test or question remain, then the possibility of unfairness is much greater. Without matching, score differences between groups reveal more about differences in attainment, not fairness.

Even with matching, the remaining differences may or may not be valid. Valid and invalid or unfair tests or questions can be distinguished by their content. If the content is important to the intended use or inference, then it should stay in the test. If the content is not important to the intended use, then the test or question should be eliminated. For example, a mathematics problem referring to something nonmathematical that is more familiar to one group than another should be eliminated.

Critical evidence for the validity and fairness of tests must come from the relationships between test scores and performance for various groups. Predictions made from SAT scores of first-year grades in college are as accurate for minority as for majority applicants. They are useful both for comparing applicants from the same ethnic group and for comparing applicants from different ethnic groups, although only scanty evidence is available for minorities other than African Americans.

Male students tend to outscore female students in quantitative areas such as mathematics and science, while female students outscore male students in verbal areas such as reading and writing. However, male test takers are not comparable to female test takers, partly because males are more likely to take high school math courses. Thus, differences in SAT mathematics scores may reflect real differences in preparation rather than gender bias in the test. The SAT predicts college grades slightly better for women than for men. But the average grades of women are paradoxically higher than the average grades of men, in spite of the women's lower average test scores. Seemingly, women tend to take courses that receive higher grades (e.g., humanities and social sciences courses), while men take courses in which fewer high grades are given (e.g., calculus and physics). Thus, the validity and fairness of a test used for prediction must be evaluated with reference to the meaning of what is being predicted.

Cole suggests that policy issues raised by test scores require more attention than they are now receiving. She believes the public should be concerned about group differences in scores on educational tests, not because of bias but because of the unequal educational opportunities that they indicate. Teachers should *not* assume that those with low scores are unable to learn. This would be a wrong inference. Rather, teachers, parents, and all citizens should be taking action to ensure that students with low scores are getting all the help they need to raise their educational performance. To infer the need for educational help is a correct inference.

Cole is also concerned about the policy implications of hugely different rates of scholarship awards to males and females, even if based on valid differences in mathematics, for example. Personally, she finds them intolerable, although from an educational perspective she understands how they occur.

We put strong requirements on tests, Cole concludes, to demonstrate validity and fairness for the inferences that are made from test scores. We should demand evidence that a test meets its intended purpose for all groups of examinees. We should also recognize that tests have been subjected to a higher standard of evidence than other measures such as grades or letters of recommendation. Just because high standards for tests focus debates and concerns about fairness on them, we should not assume that other measures will be fairer. They will not. We need to have the same concerns and validation requirements for any measures that supplement or substitute for test scores in important decisions.

Dr. Lloyd Bond, Professor, School of Education, University of North Carolina at Greensboro

Bond extends the information in the background paper by characterizing the nature of test bias and describing the techniques used to detect bias. A biased test, he says, is one that measures different attributes depending upon the subpopulation; or some of the items work to the disadvantage of particular subpopulations of examinees; or if predictions from test scores systematically over or underpredict performance for one or more subgroups of examinees.

Bond agrees that group differences in test scores are not sufficient for showing bias. Thus, the concepts of adverse impact and bias are different. He suggests that differences in test scores may reflect real differences in achievement, particularly on the mathematics section of the SAT. He would attribute the differences to different instruction and believes that tests should reflect that some children have had less favorable backgrounds.

Biased items should be eliminated from tests, but items should not be eliminated simply because they produce adverse impact.

Concerning internal validity, Bond describes various statistical approaches used to detect biased items. All of them assume the test is valid in general for all groups of examinees. If a test is categorically biased against certain groups, then any kind of internal analysis, like equating for performance on the other items as with Differential Item Functioning (the DIF method), will not get at the bias. A major shortcoming of DIF is that it flags items that genuinely distinguish between high and low scores on the test as possibly biased.

Using DIF analyses does not appear to reduce group differences very much. Bond recommends two newer, very technical approaches, the Mantel-Haenszel procedure and approaches based upon Item Response Theory (IRT), but they must be used with the item's correlation with all of the other items on the test in deciding whether to eliminate items from a test. However, even the better statistical procedures may only identify 5 to 10 percent of trial items as potentially biased.

Concerning external validity, Bond believes even very low predictive validity may still be useful for some purposes, but predictive validity alone is not sufficient for validating a test. Tests do not appear to have different predictive validity for blacks, whites, males, and females.

Bond agrees that the distinction between ability and achievement is muddled.

Bond also discusses various statistical rules used for selection and illustrates the errors that are made when tests are biased. The psychometrically sound model of fair selection results in very few minority applicants being hired, given their percentage of the applicant pool. A number of these selection models were proposed to increase the minority applicants being hired, but most result in different passing scores being used for the majority and minority groups.

Testing, Bond says, is part of our culture. He finds offensive the notion that African Americans cannot respond to that culture as well as others.

Alexandra K. Wigdor, Study Director, National Research Council, National Academy of Sciences

Wigdor summarized the results of the National Academy of Sciences study of the Department of Labor's job referral test, the General Aptitude Test Battery (GATB). When the Department of Labor expanded the GATB's use from 500 jobs to 12,000 jobs, it introduced a controversial within-group scoring system that computes the scores of African Americans, Hispanics, and all others separately according to their own group. The study addressed three issues: How valid is the GATB? Can the GATB be used for the 12,000 jobs rather than just 500? Are within-group score adjustments fair and efficient for selecting the work force?

The study concluded that the GATB makes useful but not perfect predictions; that its validity would hold for a great many jobs in the U.S. economy, particularly for the jobs for which the GATB is used; and that within-group score adjustments can be justified with the fact that errors of prediction differ for the groups. The study did not recommend proportional referral of African Americans or Hispanics, but rather making score adjustments commensurate with the prediction error so that qualified people in all groups have the same probability of being referred.

Wigdor believes the recommended score adjustment is a policy recommendation to accommodate two social goals: optimizing productivity and providing minorities with better job opportunities. But, she believes, this solution is less political because the adjustment can be justified by differences in errors that occur for high and low scores rather than by race or ethnic group *per se* (although the score is still adjusted according to one's racial or ethnic group).

The Commission also heard from two attorneys on the subject of test validity. They were both concerned primarily with recent Supreme Court decisions and the shifting of the burdens of production and proof and evidentiary standards in disparate impact cases. The Court changed the evidentiary standards by requiring the plaintiff to identify the specific selection procedure that causes the discrimination and by altering the language for the relationship the employer must demonstrate between the job and the test or other selection device.

Barry L. Goldstein, Attorney, Saperstein, Mayeda, Larkin & Goldstein

Goldstein believes that selection practices maintain job segregation. He endorses the use of tests or other screening devices when they are a business necessity but he does not support the "unfettered" use of testing or other artificial qualifications. He believes the 1971 Supreme Court decision in *Griggs v. Duke Power Co.*, and the ensuing Uniform Guidelines on Employment Selection Procedures, have been very beneficial. The caliber of employment practices has improved. Furthermore, more minorities are employed, and in better paying jobs, since Title VII of the 1964 Civil Rights Act was passed than before. Such evidence disputes claims that test score differences affect productivity and endanger the United States' competitive position in the world economy.

Goldstein believes the Supreme Court decision in *Wards Cove Packing Co. v. Atonio* will reverse the progress in civil rights because it says that showing an adverse impact is not enough to shift the burden of proof to the employer. He believes private attorneys, who litigate almost all fair employment cases, will lose unless the cases involve intentional discrimination. Thus, the *Wards Cove* decision will allow many selection procedures having adverse impact to remain in place simply because the employers did not intend to discriminate.

Goldstein objects to three aspects of the *Wards Cove* decision: (1) the requirement to pinpoint the practice that disproportionately limits opportunities; (2) the shifting of the burden of proof to the plaintiff rather than the employer; and (3) the general weakening and ambiguity of the standard of proof.

Showing that a selection practice has an adverse impact in comparison to the *relevant qualified* labor pool, Goldstein believes, should be sufficient for a *prima facie* case of disparate impact. Requiring plaintiffs also to pinpoint the objectionable selection procedure exonerates practices that create barriers in combination with other procedures. Attorneys will be too uncertain of whether the employer has maintained adequate records or of what an analysis of those records will show to risk taking cases.

Goldstein argues that it is more practical for employers to bear the burden of proof. An employer should be required to demonstrate a business necessity for a selection procedure that has adverse impact because: (1) He is responsible for choosing the selection procedure and has access to the information about it and the job for which it is used. (2) Proving a practice is justified by business necessity is easier than proving it is not. (3) The employer has more experience and resources to establish proof than the plaintiff.

Finally, the *Wards Cove* standard of proof replaces the required "business necessity" with "a reasoned review of the employer's justification for his use of the challenged practice." The former has been clarified with 20 years of litigation; the latter is unclear. It is more than a "mere insubstantial justification," but less than "essential" or "indispensable." Because the *Wards Cove* decision makes the standard of proof unpredictable, attorneys will be unwilling to represent alleged victims of discrimination.

Goldstein concludes that the pre-*Wards Cove* standards defining fair employment law should be reestablished. Removal of unnecessary barriers to employment opportunities will help accommodate the anticipated changes in demographics between now and the year 2000.

In addition, he believes selection practices can be gerrymandered to get the desirable result and have it appear neutral; employers and educators need examples of good selection practices, if not a Good Housekeeping Seal of Approval on tests; small differences in test scores are not important—a cut point should not be used; minorities should not be excluded because tests are convenient or save money; and coaching courses can boost scores, but many minorities do not have access to them.

Clint Bolick, Director, Landmark Center for Civil Rights

Bolick takes a very different position from Goldstein. Prior to the recent Supreme Court decisions, tests were automatically abandoned or invalidated, even though they may not have been discriminatory. The burdens of proof made it relatively easy to challenge tests, but nearly impossible to defend them.

He argues that in writing Title VII, Congress did not intend preferential treatment of any groups or individuals or the establishment of quotas. Their object was to permit the use of professionally developed ability tests when such tests are not designed, intended, or used to discriminate. Indeed, Bolick believes that tests help avoid discrimination because they treat all individuals the same.

Bolick agrees that the pre-*Griggs* "disparate treatment" standard, requiring plaintiffs to demonstrate that persons of different races are treated differently, was insufficient to ferret out covert instances of discrimination. Thus the "adverse impact" analysis was needed. But many explanations other than discrimination may account for disparities between the racial or ethnic composition of the community labor pool and the work force (e.g., individual preferences, qualifications, accessibility). The application of adverse impact to uncover hidden discriminatory practices was expanded to hold employers liable for discrimination whenever they used employment criteria that produced even these innocent statistical disparities.

Griggs required the employer to show his selection practices were a business necessity. The EEOC guidelines further required the employer to show that no alternative selection device is available producing less adverse impact. Lower courts changed the standard from "business necessity" to "essential" or "indispensable," a rigid standard that requires prohibitively expensive and near impossible validation by test experts. Attempts to advance an antistandards and proracial quota agenda are also evident in the National Academy of Science's study recommending unlawful race-based adjustments to GATB test scores despite the test's good predictive ability and absence of bias; and in California's ban on African Americans taking I.Q. tests which is now lifted.

This expansion, Bolick suggests, has led many employers to abandon tests and adopt quota systems at tremendous cost to our nation in productivity and competitiveness and with no assurance of solving minority problems. In fact, the suggestion that African Americans should not be held to the same standards as other people endorses a notion of their inherent inferiority.

He believes the *Wards Cove* decision "harmonized" adverse impact with the purpose of Title VII. For a *prima facie* case, plaintiffs must present statistics on the pool of qualified job applicants or the qualified population in the labor force; showing minorities are in one category of jobs and whites in another is not sufficient. Plaintiffs must focus on the specific employment practice that is potentially discriminatory; employers should not be forced to defend every employment practice and cannot be liable for the myriad of innocent causes that produce statistical imbalances in the composition of their work forces. Finally, it restores the employer's burden to proving the practice is a "business necessity" rather than essential or indispensable. Bolick believes this leaves adverse impact intact as a method of proving discrimination, but it requires that the statistics presented raise a plausible inference of discrimination.

Finally, Bolick is concerned about occupational licensing, which is regulated by States and frequently uses tests for certification or for enrollment in required curricula. Such requirements may unnecessarily restrict the trades or professions individuals pursue.

Bolick contends that with shortages of skilled labor, affirmative action solutions should focus on investing in human capital development and on expanding economic mobility, rather than quotas.

Areas of Agreement and Disagreement

The Commission's consultation on test construction issues and the longer papers supplied by the panelists convey the nature of the controversy. D. Monty Neill, writing on behalf of FairTest, and James Loewen and Barry Goldstein view testing as an obstacle to the important goals of enhancing equity and increasing opportunities. Although Nancy Cole, Lloyd Bond, and Clint Bolick also do not want tests to be unfair obstacles to opportunities, they believe that tests are merely an indicator of other inequalities that minorities face, particularly in the education they receive. They emphasize the importance of having accurate assessments because of the many different needs that tests fill.

Despite the wide range of views these panelists hold, they reveal many areas of agreement. The following section identifies some major areas of agreement and disagreement.

Definitions of Bias and Discrimination. All of the panelists recognized the potential for bias in tests and for the misuse of test scores in ways that are biased and unfair.

The testing experts agreed that average group differences in test scores alone are not evidence of bias. The attorneys also agree that such discrepancies, in and of themselves, are not proof of discrimination.

Each of the panelists listed a variety of potential causes of adverse impact. Most named differences in the quality of education.

Internal Validation—Methods for Eliminating Item Bias. All of the panelists agree that any items that are biased should be eliminated from tests, although what they regard as "biased" differs.

Although experts' judgments of test questions on their face (i.e., face validity) may be useful for eliminating offensive items, the panelists argue they are insufficient for eliminating biased items.

The panelists agree that test validation procedures must examine individual test items for bias using comparisons of statistics for relevant groups. They sharply disagree over which method should be used. Their discussion, however, suggests that some methods of comparing item statistics across groups will identify a larger proportion of potentially biased items than others. Seemingly the least stringent of these methods compares the difficulties of items across racial or ethnic groups among test takers who have similar overall test scores. Both Loewen and Neill dismiss this method as useless for identifying biased items. According to Loewen, it identifies items that *reduce* group differences in test scores as often as items that *create* them. According to Neill, it is based on circular reasoning that must first address the most important question, whether the test as a whole is biased. Nonetheless, this method identifies the minimum proportion of test items that any of the panelists believe should be examined more carefully for test biases. Thus, all agree that the method should identify at least as many items for further scrutiny as if the group comparisons were made adjusting for overall test score.

For the most part, the testing experts did not single out any of the methods that adjust for overall test score as more or less adequate than any others. If they approve of such methods, all of them are acceptable; if they believe such methods are inadequate, all of them are inadequate. Bond, however, prefers newer approaches like the Mantel-Haenszel procedure and those based upon Item Response Theory to DIF, but believes they should be used in combination with the item's correlation with all of the other items on the test in deciding whether to eliminate items from a test.

Loewen believes much more stringent statistical methods of identifying biased items should be used and suggests two methods. The one compares the difficulty of items across groups regardless of overall test score and is much simpler than the above method that adjusts for overall test score; the other involves assessing items' relationships with output variables, such as first-year college grades. Loewen does not believe that revising test items with the aid of output variables will remove all adverse impact. Cole asserts that the method is impractical.

Once a method has identified items that may be biased, opinions differ on whether or not those items must be eliminated. Although Loewen agrees that items on which groups differ in performance are not necessarily biased, he believes they should be eliminated from tests to enhance equality. Other panelists would not agree to eliminate the items these methods identify, but they may agree that test developers should provide written justification for continuing to include such items.

Extent of Bias in Existing Tests. Seldom do allegations that tests are biased quantify the extent of that bias. When they do, the extent of bias is typically characterized in one of two ways: the number of test items that are biased and the proportion of group differences in test scores due to bias. Also, attempts to quantify the extent of bias in tests have often focused on the SAT, as did these experts.

Bond estimated the number of test items that are biased by the number of items the statistical procedures identify. Even the better statistical procedures, he said, may only identify 5 to 10 percent of trial items as potentially biased. However, not all of the items identified by these methods would be considered biased, and those that were would be eliminated from the test.

Despite their different opinions about test bias and adverse impact, Bond and Loewen both concluded that the largest part of group differences on the math section of the SAT are not due to bias. Bias accounts for at most one-third of the black-white difference in math scores. Their conclusions about bias in the verbal section of the test were much less certain, although both seemed to feel that more of the difference in the verbal was due to bias than in the math.

Loewen suggests that the entire male-female difference and much of the racial group difference in average verbal scores are due to bias because of how the content domain is defined for such tests. In math, the set of fundamental operations and problems that must be mastered is finite, though large. The set of vocabulary words, analogies, contextual meanings, etc., is infinite. A consensus on what part of this verbal material is fundamental might avoid charges of test bias. However, no such agreement exists. Thus, with different groups having different exposures to such materials, test developers can manipulate group differences in scores and, Loewen suggests, construct tests with any desired difference between groups.

Methods for External Validation. Our testing experts agree that methods for eliminating item bias may not be effective when systematic biases run through all the items of a test. Thus, collecting information about how test scores relate to some criterion other than the test itself is critical for validation. Neill points out, and the others would agree, that the external criterion should not be just another test.

The panelists disagreed about whether the predictive validity of tests is the same across sexes or racial groups. They also disagreed about whether small correlations between test scores and performance were adequate for validation. However, all felt that something more than predictive validity is required for validation.

Panelists with generally opposing viewpoints agree that content should be a driving force in validation studies. For example, Neill suggests that school curricula or job duties should determine test content in education and employment applications. Cole believes that even items showing adverse impact should be included if they represent appropriate content.

All the panelists feel that more basic research is needed to understand what it is that tests measure, e.g., whether it is ability, achievement, a single dimension of intelligence, multiple facets of knowledge, learning, thinking, or problem solving. They disagree, however, about how much such research is needed before tests are useful. What responsibilities test companies or administrators share for conducting basic research in the course of test development or selection is unclear.

Monitoring of Test Construction and Use. Who sets the standards for test development and use? All panelists voiced support for some form of public involvement. Cole believes that through advisory boards and forums the public should be involved in determining what actions will be taken based upon test scores. Bond and Goldstein argue that the courts should decide policy issues. Neill and Loewen say there should be Federal oversight for test development and use. The suggestion of establishing Federal oversight for the testing industry, notably, did not draw any strong objections.

All of the experts agree that properly designed tests can be used inappropriately, in ways that bias the interpretations made of test scores. However, none speculated on how frequently inappropriate use may occur.

They all agree that important decisions, such as denial of scholarships, college admissions, or jobs, should not be based solely on test scores. Experience and education or other important selection criteria should be used too.

Mechanisms for Handling Group Differences in Test Scores. The panelists agree that issues of fairness are separate from issues of bias or adverse impact. They generally agree that adverse impact will remain in tests even if all bias is removed. However, each proposes a different solution.

Neill suggests doing away with tests in favor of "authentic" assessments such as work samples; at the very least, test scores should be only one of multiple criteria. Loewen recommends removing items showing adverse impact from tests during test construction, even if these items are unbiased. Wigdor and the National Academy of Science's report propose adjusting test scores for racial/ethnic groups by the amount of error in the test's predictions, so that successful workers in each racial/ethnic group have the same probability of being referred for the job. Her solution is milder than the Employment Service's now illegal race norming, which adjusted for the entire difference between groups, not a part of it. In discrimination cases, Goldstein would challenge employers to defend all of their selection procedures as essential for the job if any of them shows adverse impact. He would also dismiss the typically low correlations between test scores and performance as too small to validate test use.

In contrast, Cole, Bond, and Bolick think that tests should be as accurate as possible, regardless of the adverse impact they show. Providing quality education for all groups, they believe, is the key to eliminating the adverse impact that tests show. Bolick would place the burden of proving discrimination on the plaintiff, lest the employer be held liable for the myriad of innocent causes, such as differences in the quality of education across groups, that may produce adverse impact.

Conclusions

Issues of the validity of employment and education tests continue to arise in Federal, State, and local courts and before Congress. The ways in which tests are used are changing in the

Federal Government and in other public and private sectors. The major conclusions of this report are given below.

- Properly designed tests can be used inappropriately, in ways that are unfair and that bias the interpretations made of test scores. Important decisions, such as denial of scholarships, college admissions, or jobs, should not be based solely on test scores.

- Average group differences in test scores alone are not evidence of bias, nor proof of discrimination. Bias, which refers to test scores that underestimate the performance of particular groups, is different from adverse impact, which refers to differences in average test scores between groups resulting from bias or a variety of other causes, such as differences in the quality of education. Adverse impact will likely remain in tests even if all bias is removed.

- Methods for eliminating item bias may not be effective when systematic biases run through all the items of a test. Therefore, collecting information about how test scores relate to criteria other than the test itself, such as job or school performance, is crucial for validation.

- Biased items should be eliminated from tests. Experts' prima facie judgments of tests questions are not adequate for identifying biased items. Test validation procedures must examine individual test items for bias using comparisons of statistics for relevant groups, for instance by comparing the difficulties of items across racial or ethnic groups among test takers who have similar overall performance. Once a method has identified items as potentially biased, test developers should provide written justification for continuing to include such items in their tests.

- Standards for test development and use should be set with some form of public involvement, whether it is through Federal oversight or public input on advisory boards and forums.

Glossary of Testing Terms

Ability Test—A test that estimates a person's current or future performance in some defined domain of cognitive, psychomotor, or physical functioning, employing items on which performance can be objectively determined to be right or wrong, better or poorer. See also aptitude test and achievement test.

Achievement Test—A test that measures the extent to which a person commands a body of information or possesses a skill, usually after training or instruction specifically intended to impart that information or skill. See also ability test and aptitude test.

Alternate Forms—Two or more tests intended to measure the same psychological dimension and having questions that are similar in number, type, content, difficulty, etc.

Aptitude Test—A test that is usually not closely related to a specific curriculum and that is used primarily to predict future performance, especially in education or a training program. Compare Achievement Test. The distinction between aptitude tests and achievement tests often depends on differences in test use rather than in test content.

Bias—See test bias.

Classification Error—(1) The proportion of inconsistent or incorrect categorizations of examinees that would be made on repeated administrations of the test, assuming no changes in the examinees' true performance levels. (2) The assignment of an examinee to the wrong category, such as passing a person who lacks minimal competence and should fail.

Classification Rates—The proportions of examinees placed in various categories, such as pass/fail, on the basis of test scores.

Competency Test—An achievement test designed to demonstrate whether a student or trainee has reached a given level of proficiency in some basic skill(s) or domain(s) of knowledge.

Construct—A psychological characteristic (writing ability, numerical ability, logical reasoning) considered to vary across individuals. A construct (for example, mental ability) is a theoretical concept that is inferred from empirical evidence (such as performance on a test) and is not directly observable.

Construct Validation—The process of establishing the meaning of a psychological attribute by using a set of lawlike statements or a chain of inference relating the construct to other constructs and facts (such as evidence of predictive or content validity).

Content Domain—A body of knowledge or set of tasks or behaviors defined so that given knowledge or behaviors may be classified as included or excluded.

Content Validation—The process of establishing that the testaccurately represents a balanced and adequate sampling of the relevant content domain and that it excludes content outside that domain.

Correct Answer (or Response) Rate—The percentage of people who give the correct answer to a test question. It is one index of item difficulty.

Correlation—An index of the degree of relationship between two variables, expressed as a number ranging from -1.00 (a perfect negative relationship, where high values of one variable are associated with low values of the other) to +1.00 (a perfect positive relationship, where high values of one variable are associated with high values of the other) with 0 representing no relationship.

Point-Biserial Correlation—A special correlation that is appropriate when one variable is dichotomous (e.g., a test question that is right or wrong) and the other is continuous (e.g., a test score that can be any number between 0 and 100).

Criterion—That which is predicted by a test. It may be a measure of academic or job performance or job behavior, such as achievement, productivity, accident rate, absenteeism, tenure, reject rate, training score, and supervisory or coworker rating.

Criterion Validity—When test scores are systematically related to one or more measures of performance (e.g., academic performance or important elements of job performance or work behavior).

Predictive Validity—A form of criterion validity where the test scores are systematically related to some *future* criterion that the test scores can thereafter predict (e.g., using test scores from high school to predict college performance). The correlation coefficient that measures the degree of the systematic relationship is called a validity coefficient.

Criterion Relevance—The extent to which the measure used in assessing a test's predictive validity is related to the test's intended purpose.

Criterion-referenced Test—An instrument for which score interpretations refer to an ability to perform certain tasks rather than to the performance of others.

Critical Score—A test's passing score, especially when it is the same for all applicant groups (compare cutoff score); a designated point in a distribution of scores at or above which candidates are considered successful.

Cultural Bias—A bias that occurs when test items contain information that is specific to the culture of one group and absent, to some degree, from the culture of another group.

Culture Reduced Test—Typically a test that is nonlanguage and nonscholastic in nature and does not call for any specific prior information other than an understanding of test instructions.

Cutoff Score—A test score below which candidates are rejected, especially when it is dependent on the number of openings and the number of applicants.

Differential Item Functioning—When item response theory identifies items that groups responded to differently, but the items will be subjected to further scrutiny before being labeled as biased because the statistical method is so extremely sensitive to such differences.

Differential Prediction—When test scores predict performance on some criterion, for example, college grades, differently (i.e., either too high or too low) for members of some subgroup than for test takers in general; in technical terms, when use of a common regression equation results in systematic nonzero errors of prediction for subgroups.

Difficulty Index—Any one of a variety of indices used to signify the difficulty of a test question. The percentage of some specified group, such as students of a given age or grade, who answer an item correctly is an example of one such index.

Distribution—See frequency distribution.

Egalitarian Assumption—An assumption that all racial/ethnic or gender groups should have the same average test score.

External Validity—When a test measures what it ought to as demonstrated by the relationship of test scores to other factors, usually performance of the sort for which the test selects.

Face Validity—The appearance that a test (or test item) measures the trait or ability that it is intended to measure, as judged by inspection of the test (or item).

Factor Analysis—A statistical procedure that clarifies the nature of the phenomena (constructs or "factors") measured by a test and identifies the test items most associated with them. For example, it indicates the items that distinguish best between high and low scorers on the test.

Frequency Distribution—A tabulation of data such as test scores from high to low showing the number of individuals who obtain each score or whose scores fall in each score interval.

Internal Validity—When a test measures what it ought to as demonstrated by the properties of the test itself, such as its item difficulties and point-biserial correlations.

Item—A test question or the subpart of a question that requires a response.

Item Analysis—A statistical procedure that determines the suitability of any specific test item for inclusion in a particular test. Data often provided are the difficulty of the question, the number of people choosing each multiple choice answer, and information on how well the item discriminated among the examinees with respect to a chosen criterion.

Item Bias—When individuals of a particular group respond correctly to a test item substantially more or less often than those of the overall population and this disparity stems from factors that the item is not intended to measure rather than from factors it is intended to measure.

Item Difficulty—See correct answer rate and difficulty index.

Item Response—(1) A person's answer to a question. (2) Performance on a test question rated as "right" or "wrong"; "better" or "worse."

Item Response Theory—A set of propositions that uses mathematical models to relate people's performance on test questions to their characteristics and to characteristics of the items. It is based on the assumption that the probability of a person's correct response to an item can be calculated from an estimate of the examinee's ability and characteristics of the item such as the item difficulty.

Item Type—The format of a test question referring, for example, to multiple-choice vs. free-response questions at one level and to questions about, say, synonyms vs. antonyms at another.

Job Analysis—A procedure undertaken to understand job duties and behaviors and performance standards for the job.

Job Description—A written statement of the results of the job analysis including job duties and activities, indications of the complexity and relative importance of the more significant duties or activities and/or work products.

Job Relatedness—The inference that scores on a selection instrument are relevant to performance or other behavior on the job; job relatedness may be demonstrated by appropriate criterion-related validity coefficients or by gathering evidence of the relevance of the content of the selection instrument, or of the construct measured.

Mean—Arithmetic average; the sum of a set of scores (or other values) divided by the number of scores (or values).

Mean Differences—Average differences between groups as in test scores or correct answer rates.

Measurement Error—The deviation of an obtained measure from the true value, where the hypothetical true value is assumed to be the mean of an infinite number of measurements of the same thing.

Median—The middle score in a distribution; the 50th percentile; the point that divides the group into two equal parts. Half of the group's scores fall below the median and half above it.

Minimum Competency Test—An achievement test designed to demonstrate whether a student or trainee has reached a minimally acceptable level of proficiency in some basic skill(s) or domain(s) of knowledge.

Mode—The score or value that occurs most frequently in a tabulation of data.

Normal Distribution—A theoretical distribution that describes the expected frequencies of most social data because it is based on the laws of probability. The graphical representation of a normal distribution is bell-shaped, high in the center, low at the ends, and perfectly symmetrical; the mean, median, and mode coincide at the center. There is a specific known equation for the normal distribution. Not all bell-shaped distributions are normal.

Norms—Descriptive statistics for well-defined groups that are logical references for other individuals who take the test.

Norm-referenced Test—An instrument for which interpretation is based on the comparison of a test taker's performance to the performance of other people in a specified group.

Overinterpretation of Test Scores—The extension of test scores from domains in which they are valid to broader areas or domains where they are not.

P-Level—See correct answer rate.

Parallel Forms—Two or more tests intended to measure the same psychological dimension and having questions that are the same in number, type, content, difficulty, etc. See alternate forms. More of the statistical and content specifications must be the same for "alternate forms" to be called "parallel."

Passing Rate—The percentage of a group scoring above a critical score.

Percentile—A point (score) in a distribution below which falls the percentage of cases indicated by the given percentile. Thus, the 15th percentile denotes the score or point below which 15 percent of the scores fall. (Also known as centile.)

Percentile Rank—The percentage of scores in a distribution equal to or lower than a particular obtained score.

Performance—The effectiveness and value of work behavior and its outcomes.

Performance Standard—A critical score or a defined level of performance on some task. For example, "Run 100 yards in 12 seconds or less."

Pilot Subsection—An unscored section of a test included to try out new items for inclusion in future tests.

Pilot Testing—Small-scale tryout of test questions or a test form, often involving observation of and interviews with examinees.

Population Subgroup—A part of the larger population that is definable according to various criteria as appropriate (e.g., by sex, race or ethnic origin, training or formal preparation, geographic location, income level, handicap, or age).

Predictive Validity—See criterion validity.

Predictor—A measure used to predict criterion performance, for example, scores on a test, or judgments of interviews.

Pretest—A test designed for the purpose of trying out new items and obtaining statistics for them before they are used in a final form.

Psychometricians—Those who engage in psychometrics.

Psychometrics—(1) The measurement of psychological characteristics such as aptitudes, personality traits, achievement, skill, and knowledge. (2) The study of properties of psychological measurements, especially tests and test items. The properties of tests may include test construction methods, speededness (see below), length, reliability, stability, validity, and bias; properties of test items may include level of difficulty, bias, distractors, and their effects.

Quartile—One of three points (scores) that divides the cases in a distribution into four equal groups. The lower quartile, or 25th percentile, sets off the lowest fourth of the group; the middle quartile, is the same as the 50th percentile, or median; and the third quartile, or 75th percentile, marks off the highest fourth.

Race-by-item Interaction—When correct response rates differ by race such that, relative to other test items, one or more items are much more difficult (or easy) for one race than for another; a difference in the rank order of *p*-levels.

Regression Equation—An algebraic equation for the best fitting line used to predict criterion performance from predictor scores.

Reliability—The extent to which a test is consistent in measuring whatever it does measure or the degree to which repeated measurement of the same individual would tend to produce the same result; consistency or dependability or repeatability.

Respondent—An individual who provides data to a research project, particularly by answering a questionnaire or taking a test. See subject.

Scaled Score—A score on a test expressed as a number or position on a standard reference scale, such as the 200 to 800 scale for College Board tests. Scores are converted to a scale so that they are independent of the particular form of the test and of the composition of the group of examinees who took it.

Score—A quantitative or categorical value (such as "pass" or "fail") assigned to an examinee as the result of some measurement procedure.

Selection Instrument—Any method or device, such as a test, used to evaluate characteristics of persons for purposes of selection.

Selection Model—A rule for arriving at a selection decision, especially when it uses test scores and uses social values to adjust for the uncertainty in them.

Skewness—Asymmetry in a distribution. If the scores tend to spread out more when the values are high, the distribution is positively skewed; if they tend to spread out more when the values are low, it is negatively skewed.

Speededness—The extent to which a test taker's score depends on the rate at which work is performed rather than on the correctness of the response. One indicator of speededness is the percentage of test takers who do not complete the test.

Speed Test—A test in which performance is measured by the number of tasks performed in a given time. Examples are tests of typing speed and reading speed. Also, a test scored for accuracy where the test taker works under time pressure.

Standard Deviation—A statistic characterizing the magnitude of the differences among a set of measurements; a measure of dispersion of a frequency distribution. It is the square root of the average squared difference between each measurement and the mean of the measurements.

Standard Score—A score that describes the location of a person's score within a set of scores in statistical terms—distance from the mean in standard deviation units.

Standardization—A test development procedure designed to distribute measured characteristics (e.g., aptitude, achievement) of the test-taking population across high and low scores with meaningful distinctions. First, items are selected that range around an appropriate level of difficulty (i.e., according to the performance of a population of test takers), then each test taker's number of correct answers is converted to scores that express his/her standing relative to others of appropriate age or (grade) level.

Standardized Predictor—A test employed for estimating a criterion of job performance, the test having been developed and standardized according to professionally prescribed methods.

Standardized Test—A test administered and scored under conditions uniform to all test takers in order to make test scores comparable and to ensure that test takers have equal chances to demonstrate what they know.

Statistical Control—A procedure that mathematically removes unwanted effects of some variables, biases or error, for better understanding the relationships between the remaining variables. The simplest form of a statistical control examines the variables of interest among individuals having the same value of the unwanted variable.

Statistical Significance—A scientific result is larger or occurs more often than one would expect by chance alone. Large numbers of observations will produce statistically significant results even though the magnitude of the result is quite small. Thus, statistically significant results may be of no practical importance. Similarly, results may be insignificant simply because the number of observations is small.

Subject—An individual who participates in a research project, particularly in a laboratory experiment. See respondent.

Test—A sample of questions or tasks from a domain that is used to make inferences about a person's, a group's, or an institution's performance.

Test Analysis—A description of the statistical characteristics of a test following administration, including but not limited to distributions of item difficulty and discrimination indices, score distributions, mean and standard deviation of scores, reliability, and indications of speededness.

Test Bias—A pattern of errors in test scores that systematically effect some groups but not others.

Test Developers—People who construct tests or who set policies for particular testing programs.

[ETS] Testing Program—A comprehensive ongoing service under which examinees are scheduled to take a test under standardized conditions, the tests are supplied with instructions for giving and taking them, and arrangements are made for scoring the tests, reporting the scores, and providing interpretative information. A program is characterized by its continuing character and by the inclusiveness of the services provided" (*ETS Standards*, 1987: 36).

Test-Retest Reliability—An estimate of reliability based on the correlation between scores on two administrations of the same test to the same group of people. See reliability.

Test Users—People who choose tests, commission test development services, or make decisions on the basis of test scores.

Trait—An enduring characteristic of a person that is common to a number of that person's activities.

True Score—The hypothetical average of the scores earned by an individual on an unlimited number of perfectly parallel forms of the same test.

Truth in Testing Movement—"A variety of efforts to regulate standardized testing, many of which have taken the form of legislative proposals to require that (a) individual test takers have access to corrected test results within a specified period after test administration; (b) test sponsors or publishers file information on test development, validity, reliability, and cost with government agencies; and (c) testing agencies give individual test takers information on the nature and intended use of tests prior to testing and guarantee their right of privacy concerning their own test scores" (Haney, 1981).

Type I Error—Concluding that a significant relationship exists when it does not.

Type II Error—Concluding that no significant relationship exists when it does.

Utility—The practical usefulness of a selection instrument that allows the user to make quick and accurate decisions that save time or money, improve efficiency, or have other beneficial effects for either the test taker or user.

Validation—The evaluation of the appropriateness and meaningfulness of interpretations from scores on a test. The process does not necessarily guarantee approval of the test, because the research may conclude the test has little validity.

Validity—The degree to which a test measures what it is supposed to measure, that is, inferences from its scores are appropriate or meaningful as supported by evidence. Three types of validity are content validity, criterion validity, and construct validity.

Validity Coefficient—A coefficient of correlation that shows the strength of the relation between predictor and criterion.

Validity Generalization—The use of results of validity studies obtained in one or more studies to justify inferences about job behavior or job performance in jobs or groups of jobs in different settings.

Variability—The spread or scatter of scores.

Variable—A quantity that may take on any one of a specified set of values.

Variance—A statistic characterizing the magnitude of the differences among a set of measurements; a measure of dispersion of a frequency distribution. It is the average squared difference between each measurement and the mean of the measurements. The square root of the variance is known as the standard deviation.

Z-Scores—Standard scores calibrated in commonly used statistical units: For the group used in defining the scale, the scores have a mean (an average) equal to zero and a standard deviation of one unit.

Table of Select Cases

Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975).

Allen v. Alabama State Bd. of Educ., 612 F. Supp. 1046 (M.D. Ala. 1985) *vacated* 636 F. Supp. 64 (1986), *rev'd* 816 F.2d 575 (11th Cir. 1987), *reh. denied* 817 F.2d 761 (1987).

Anderson v. Banks, 520 F. Supp. 472 (S.D. Ga. 1981).

Brown v. Bd. of Educ., 347 U.S. 483 (1954).

Crawford v. Honig, No. C–89–0014–RFP (N.D.Cal. May 10, 1988).

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979).

Golden Rule Insurance Co. v. Washburn, 1984, (No. 419–76, Ill. 7th Jud. Cir.).

Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Guardian Ass. v. Civil Service Comm. of New York, 463 U.S. 582 (1983).

Hobsen v. Hansen, 269 F. Supp. 401 (D.D.C. 1967), *cert. dismissed* 393 U.S. 801 (1968).

Lorance v. AT&T Technologies, Inc., 490 U.S. 900 (1989).

Larry P. v. Riles, 495 F.Supp. 926 (N.D. Cal. 1979) *aff'd in part and rev'd in part* 793 F.2d 969 (9th Cir. 1984).

Luevano v. Campbell, 93 F.R.D. 68 (D.D.C. 1981).

Montgomery v. Starkville Muni. Separate School Dist., 854 F.2d 127 (5th Cir. 1988).

Morey v. Doud, 354 U.S. 457 (1957).

Palmer v. Shultz, 616 F. Supp. 1540 (D.D.C. 1985), *rev'd and remanded*, 815 F.2d 84 (D.C. Cir. 1987).

Parents in Action on Special Education (PASE) v. Hannon, 506 F. Supp. 831 (D.C. Ill. 1980).

Patterson v. McLean, 491 U.S. 164 (1989).

Quarles v. Oxford Muni. Separate School Dist., 868 F.2d 750 (5th Cir. 1989).

Sharif by Salahuddin v. New York State Educ. Dept., 709 F. Supp. 345 (S.D. N.Y. 1989).

LULAC v. United States, 793 F.2d 636 (5th Cir. 1986).

Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989).

Watson v. Fort Worth Bank, 487 U.S., 977 (1988)(Plurality opinion).

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing* (Washington, DC: American Psychological Association, 1985).

American Psychological Association, *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (Washington, DC: 1954).

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Tests and Manuals* (Washington, DC: 1966).

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Tests* (Washington, DC: 1974).

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, "Guidelines for Nonsexist Language in APA Journals," pp. 43-49 in *Publication Manual of the American Psychological Association* (3rd edition), (Washington, DC: 1986); or *American Psychologist*, 1977, 32(6), 487-494.

Anrig, Gregory R., "ETS on 'Golden Rule'," *Educational Measurement: Issues and Practice*, Fall 1987: 24-27.

Bradley, Ann, "Concern Voiced Over National Teacher Certification," *Education Week*, Feb. 21, 1990, p. 8.

Cherryholmes, Cleo H., "Construct Validity and the Discourses of Research," *American Journal of Education*, May 1988: 421-456.

Cole, Nancy, "The Implications of Coaching for Ability Testing," pp. 389-414 in Wigdor, Alexandra K. and Wendell R. Garner, *Ability Testing: Uses, Consequences, and Controversies*, Part II: Documentation Section (Washington, DC: National Academy Press, 1982).

The College Board, "Guidelines on the Uses of College Board Test Scores and Related Data" (New York: 1988).

Delahunty, Robert J., "Perspectives on Within-Group Scoring," *Journal of Vocational Behavior*, December 1988, 33(3): 463-477.

Dwyer, Carol A., "Test Content and Sex Differences in Reading," *The Reading Teacher*, May 1976: 753-757.

Education Daily, "ETS to Release New Teacher Licensing Tests in Fall," Mar. 29, 1990, p. 4.

Education Daily, "Bush Reveals Master Plan for Education Overhaul By 2000," Apr. 19, 1991, pp. 1-3.

Educational Testing Service, *ETS standards for quality and fairness* (Princeton, NJ: 1987).

Educational Testing Service, "Guidelines for Proper Use of NTE Tests" (Princeton, NJ: 1985).

Elliott, Rogers, *Litigating Intelligence: IQ Tests, Special Education, and Social Science in the Courtroom* (Dover, MA: Auburn House Publishing Company, 1987).

Elliott, Rogers, "Tests, Abilities, Race and Conflict," *Intelligence*, 1988, 12: 333-350.

Elliott, Rogers, and A. Christopher Strenta, "Effects of Improving the Reliability of the GPA on Prediction Generally and on Comparative Predictions for Gender and Race Particularly," *Journal of Educational Measurement*, Winter 1988, 25(4): 333-337.

Equal Employment Opportunity Commission, "Guidelines on Employee Selection Procedures" 29 C.F.R. Part 1607 (1966).

Equal Employment Opportunity Commission, "Guidelines on Employee Selection Procedures," 35 *Fed. Reg.*, 12333-36 (1970).

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, "Uniform Guidelines on Employee Selection Procedures (1978)," 29 CFR Part 1607 (1991); 8 FEP (BNA) § 401:2231-2272.

Equal Employment Opportunity Commission, Office of PersonnelManagement, Department of Justice, Department of Labor and Department of the Treasury, "Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures," 8 FEP (BNA) § 401:2301-2329.

Evangelauf, Jean, "Critics and Defenders of Admission Tests Eye Court's Limit on Use," *Chronicle of Higher Education*, v. 35, no. 23, Feb. 15, 1989, pp. A1, 32.

FairTest (National Center for Fair & Open Testing), "FairTest Examiner," Spring 1987, 1(1): 4.

Fields, Cheryl M., "Close to 100 Pct. of Grambling U. Students Now Pass Teacher-Certification Examination, Up From 10 Pct.," *The Chronicle of Higher Education*, Nov. 23, 1988a, pp. A23–A24.

Fields, Cheryl M., "Critics Question Validity of Standardized Tests for Would-Be Teachers," *The Chronicle of Higher Education*, Nov. 30, 1988b, pp. A33–A34.

Fiske, Edward B., "Disputed Exam Teachers Take Being Replaced," *New York Times*, Friday, Oct. 26, 1988, pp. A1, B5.

Flaugher, Ronald L., "The Many Definitions of Test Bias," *American Psychologist*, July 1978: 671–678.

Glaberson, William, "Scholarships Based on S.A.T.'s are Unfair to Girls, Court Rules," *New York Times*, Feb. 4, 1989, p. 1.

Gold, Michael E., "Griggs' Folly: An Essay on the Theory, Problems, and Origins of the Adverse Impact Definition of Employment Discrimination and a Reconsideration for Reform," *7 Indus. Relations L.J.* 429–598 (1985).

Goldstein, Amy, "Half of black applicants fail teaching test," *The Baltimore Sun*, Sept. 13, 1987, p. 1A.

Gonzales, Enrique J., "Lawmaker charges discrimination at State," *The Washington Times*, Oct. 13, 1989.

Gordon, Robert A., Mary A. Lewis, and Ann M. Quigley, "Can We Count on Muddling through the *g* Crisis in Employment," *Journal of Vocational Behavior*, December 1988, 33(3): 424–451.

Gottfredson, Linda S. (Ed.), "The *g* Factor in Employment" [Special Issue], *Journal of Vocational Behavior*, December 1986, 29(3): 293–461.

Gottfredson, Linda S., "Reconsidering Fairness: A Matter of Social and Ethical Priorities," *Journal of Vocational Behavior*, December 1988, 33(3): 293–319.

Gottfredson, Linda S., and James C. Sharf (eds.), "Fairness in Employment Testing" [Special Issue], *Journal of Vocational Behavior*, December 1988, 33(3): 225–490.

Haney, Walt, "Validity, Vaudeville, and Values," *American Psychologist*, October 1981, 36(10): 1021–34.

Hartigan, John A., and Alexandra K. Wigdor, (Eds.), *Fairness in Employment Testing* (Washington, DC: National Academy Press, 1989)

Havemann, Judith, "New Hiring Test Called Fair to Minorities," *The Washington Post*, June 24, 1988.

Havemann, Judith, "Fewer Take New Test for Federal Jobs," *The Washington Post*, July 11, 1990, p. A7.

Holden, Constance, "Court Ruling Rekindles Controversy Over SATs," *Science*, v. 243, Feb. 17, 1989, pp. 885–887.

Jensen, Arthur R., *Bias in Mental Testing* (New York: The Free Press, 1980).

Joint Committee on Testing Practices, *Code of Fair Testing Practices in Education* (Washington, DC: American Psychological Association, 1988).

Linn, Robert, "Issues of validity for criterion-referenced measures," *Applied Psychological Measurement*, 1980, 4: 547–561.

Linn, Robert, "Ability Testing: Individual Differences, Prediction and Differential Prediction," pp. 335–388 in Alexandra K. Wigdor and Wendell R. Garner (Eds.), *Ability Testing: Uses, Consequences, and Controversies* (Washington, DC: National Academy Press, 1982).

Madaus, George, and Benjamin Shimberg, "Assuring the Quality and Fairness of 'High Stakes' Tests," unpublished manuscript, Feb. 28, 1989.

Merl, Jean, "Court Ban on IQ tests for Blacks Sparks Parents' Suit," *The Los Angeles Times*, Aug. 5, 1991.

North Carolina Advisory Committee to the United States Commission on Civil Rights, *In-School Segregation in North Carolina Public Schools,*" March 1991.

Novick, Melvin R., "Federal Guidelines and Professional Standards," *American Psychologist*, October 1981, 36(10): 1035–46.

Pennock-Roman, Maria, "The Status of Research on the Scholastic Aptitude Test (SAT) and Hispanic Students in Post-secondary Education," in Bernard R. Gifford and Linda C. Wing (eds.), *Current Views on Testing in Education: A Policy Perspective* (Boston, MA: Kluwer Academic Publishers, 1991).

Professional Regulation News, "Rhode Island Assembly Approves Teachers Examination Moratorium," v. 9, no. 10, October 1989, p. 2.

Purnell, John, "Bias ruling shelves Foreign Service test results," *The Washington Times*, Mar. 9, 1989.

Reilly, Richard R., and Michael A. Warech, "The Validity and Fairness of Alternative Predictors of Occupational Performance," in Linda C. Wing and Bernard R. Gifford (eds.), *Employment Testing: Linking Policy and Practice* (Boston, MA: Kluwer Academic Publishers, 1991).

Rose, David L., "Subjective Employment Practices: Does the Discriminatory Impact Analysis Apply?" *25 San Diego L. Rev.*, 63–93 (1988).

Rothman, Robert, "Promise, Pitfalls Seen in Creating National Exams," *Education Week*, v. 10, no. 19, Jan. 30, 1991, pp. 1, 18.

Schmidt, Frank L., "The Problem of Group Differences in Ability Test Scores in Employment Selection," *Journal of Vocational Behavior*, 1988, 33: 272–292.

Schmidt, Frank L., and John E. Hunter, "Racial and Ethnic Bias in Psychological Tests: Divergent Implications of Two Definitions of Test Bias," *American Psychologist*, January 1974: 1–8.

Schmidt, Frank L., and John E. Hunter, "Employment Testing: Old Theories and New Research Findings," *American Psychologist*, 1981, 36(10): 1128–37.

Seymour, Richard T., "Why Plaintiffs' Counsel Challenge Tests, and How They Can Successfully Challenge the Theory of 'Validity Generalization'," *Journal of Vocational Behavior*, 1988, 33: 331–364.

Shepard, Lorrie A., "The Case for Bias in Tests of Achievement and Scholastic Aptitude," pp. 177–190 in Modgil, Sohan, and Celia Modgil in *Arthur Jensen: Consensus and Controversy* (New York: The Falmer Press, 1987).

Shimberg, Benjamin, *Occupational Licensing: A Public Perspective* (Princeton, NJ: Educational Testing Service, 1982).

Snyderman, Mark and Stanley Rothman, *The IQ Controversy, the Media and Public Policy* (New Brunswick, NJ: Transaction Books, 1988).

Society for Industrial and Organizational Psychology, *Principles for the validation and use of personnel selection procedures* (third edition) (College Park, MD: 1987).

Steele, Shelby, "The Recoloring of Campus Life," *Harper's Magazine*, Feb. 1989: 47–55.

Strenta, A. Christopher, and Rogers Elliott, "Differential Grading Standards Revisited," *Journal of Educational Measurement*, Winter 1987, 24(4): 281–291.

Swoboda, Frank, and Judith Havemann, "Labor Dept. Abandoning Blue-Collar Aptitude Test," *The Washington Post*, July 11, 1990.

Tittle, Carol K., Karen McCarthy and J. F Streckler, *Women and Educational Testing* (Princeton, NJ: Educational Testing Service, 1974)

Uhlig, Mark A., "New York State Returns to National Tests for Scholarships," *New York Times*, Oct. 6, 1989, p. B1.

U.S. Commission on Civil Rights, *Report of the United States Commission on Civil Rights On the Civil Rights Act of 1990* (Washington, DC: July 1990)

U.S. Department of Education, "America 2000: An Education Strategy," 1991.

U.S. Department of Labor, Employment and Training Administration, News Release (USDL: 90–354), "Dole Suspends Use of Job Aptitude Test."

U.S. Department of Labor, Employment and Training Administration, "Proposed Revised Policy on Use of Validity Generalization—General Aptitude Test Battery for Selection and Referral in Employment and Training Programs; Notice and Request for Comments," *Federal Register*, vol. 55, no. 142, July 24, 1990, pp. 30162–64.

U.S. General Accounting Office, "Minorities and Women are Underrepresented in the Foreign Service," June 1989.

Vukelich, Dan, "OPM behind schedule on new recruitment plan," *The Washington Times*, May 25, 1989.

Watkins, Beverly T., "New Tests Expected to Bring DramaticChanges in the Way Prospective Teachers Are Assessed," *The Chronicle of Higher Education*, Nov. 9, 1988, pp. A1, A36.

Wigdor, Alexandra K., and Wendell R. Garner, *Ability Testing: Uses, Consequences, and Controversies*, Part I: "Report of the Committee" (Washington, DC: National Academy Press, 1982a).

Wigdor, Alexandra K., and Wendell R. Garner, *Ability Testing: Uses, Consequences, and Controversies*, Part II: Documentation Section (Washington, DC: National Academy Press, 1982b).

Wigdor, Alexandra K. and John A. Hartigan (eds.), *Interim Report: Within-Group Scoring of the General Aptitude Test Battery* (Washington, DC: National Academy Press, 1988).

Appendix A

Federal Guidelines and Professional and Agency Standards

Principles, guidelines, standards, and a code have been issued by Federal agencies, professional associations, and test developers to protect test takers and ensure the quality of tests and their usage. Protection is necessary because the "interests of the various parties in the testing process are sometimes congruent and sometimes not" (Novick, 1981).

The American Psychological Association (APA) was first to issue guidelines. Other agencies and organizations have based their guidelines and principles on the APA *Standards*. The "Uniform Guidelines on Employee Selection Procedures," the *Principles for the Validation and Use of Personnel Selection Procedures*, and the "Code of Fair Testing Practices in Education" were designed to be consistent with the APA *Standards*.

Although each document has a special audience and purpose (see below), the "Uniform Guidelines" differ in two important ways. First, because they are published in the *Code of Federal Regulations*, the "Uniform Guidelines" assist employers and others in complying with the requirements of Federal law prohibiting unlawful employment practices. The "Guidelines" also provide a framework for determining the proper legal use of tests and other selection procedures. None of the other documents has legal standing. Second, the APA *Standards* and the "Uniform Guidelines" interpret the same standards as though they occur at different levels. The APA *Standards* were intended as ideals toward which professionals should strive in validating tests. Courts have interpreted the "Uniform Guidelines" as establishing minimum requirements for test validation, though the "Guidelines" do not require test validation of selection procedures where no adverse impact results (29 C.F.R. 1607.1(B)). As minimum requirements, the standards for test validation are not affordable or achievable for many employers. Many have called for revisions of the "Uniform Guidelines" because of this discrepancy. The Equal Employment Opportunity Commission has promised a revision, but has yet to report any progress in this direction.

Except for the "Uniform Guidelines," none of the associations or organizations issuing standards, principles, and guidelines has any means of enforcing them. A national organization or agency designed to monitor and regulate testing has often been proposed. The Ford Foundation is currently funding a study of the feasibility of establishing such an agency.

Test developers and testing programs may develop their own monitoring systems. One well-known test developer—Educational Testing Service (ETS)—issues agency *Standards* and monitors adherence both internally and through an annual review conducted by a Visiting Committee of persons outside the agency. The *ETS Standards*, the College Board's *Guidelines*, and the National Teacher Examinations (NTE) "Guidelines" are also described below. They are examples of ETS's agency standards and those developed with test users for two widespread testing programs.

Standards for Educational and Psychological Testing

Short title: APA *Standards*

Developed by: Joint Committee of the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME)

Publication Date: 1985

Precursors: Standards for Educational and Psychological Tests (1974), *Standards for Educational and Psychological Tests and Manuals* (1966), and *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (1954)

Purpose: To provide a set of technical guidelines for the evaluation of tests, testing practices, and the effects of test use. The Standards represent evolving ideals towards which professionals should strive rather than a prescriptive check list of minimum standards. Thus, in evaluating the acceptability of a test or its application, these standards should be used along with professional judgment based on a knowledge of behavioral science, psychometrics, and the field to which the tests apply and the availability and feasibility of alternatives.

Scope: All testing situations, including clinical testing, educational testing, psychological testing in the schools, test use in counseling, employment testing, professional and occupational licensure and certification, and program evaluation.

Intended Users: Test developers, test users (e.g., employers, counselors), and test administrators.

Topics Covered: Test construction, evaluation, scoring and administration, the rights of test takers, and special concerns with linguistic minorities and those with handicapping conditions. They discuss validity in depth, paying particular attention to construct-, content-, and criterion-related evidence; validity generalization; and differential prediction. Standards are developed for each test application (e.g., clinical testing, educational testing, program evaluation). Standards to protect the rights of test takers recommend, for example, only authorized disclosure of test results, the avoidance of stigmatizing labels based upon test results, and methods of handling testing irregularities such as misconduct. They do not address "truth in testing" issues.

Reference: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, (Washington, DC: American Psychological Association, 1985).

Copies Available From: American Psychological Association, Inc., 1200 Seventeenth Street, N.W., Washington, D.C. 20036

"Uniform Guidelines on Employee Selection Procedures (1978)"

Short Title: "Uniform Guidelines"

Developed By: Equal Employment Opportunity Commission (EEOC), Civil Service Commission, Department of Labor, and Department of Justice

Publication Date: 1978

Precursors: EEOC's "Guidelines on Employee Selection Procedures" (1970) and (1966)

Purpose: To establish a uniform set of principles on selection procedures and the proper use of tests, and to aid compliance with the requirements of Federal law prohibiting employment practices that discriminate on grounds of race, color, religion, sex, and national origin. The *Uniform Guidelines* were intended to be consistent with the *APA Standards* issued in 1974.

Scope: Employment testing, especially where an adverse impact occurs, including tests used for hiring, promotion, demotion, membership (for example, in a labor organization), referral, retention, and licensing and certification.

Intended Users: Employers, labor organizations, and employment agencies, and licensing and certification boards.

Topics Covered: Definitions of discrimination and adverse impact; standards for validity studies, including acceptable types of validity studies (e.g., criterion, content, and construct validity), the choice of criterion measures, the adequacy of research methodology and size of statistical relationships in predictive validity studies, job analysis as a requirement for content or construct validity, the use of cutoff scores; generalizing validity studies across jobs and employers and across races, sexes, and ethnic groups; fairness; and the policy of affirmative action.

Reference: Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice, "Uniform Guidelines on Employee Selection Procedures (1978)," 29 C.F.R. Part 1607 (1991). For the "Guidelines" with legislative history included as introductory material, see 8 FEP (BNA), § 401:2231-72. For the full text of 90 interpretive "Questions & Answers" on the guidelines, see: Equal Employment Opportunity Commission, Office of Personnel Management, Department of Justice, Department of Labor, and Department of the Treasury, "Adoption of Questions and Answers to Clarify and Provide a Common Interpretation of the Uniform Guidelines on Employee Selection Procedures," 8 FEP (BNA) § 401:2301-29.

Copies Available From: The Bureau of National Affairs, Inc., 2445M Street N.W., Suite 275, Washington, D.C. 20037

Principles for the Validation and Use of Personnel Selection Procedures

Short Title: (Division 14) *Principles*

Developed By: Society for Industrial and Organizational Psychology (Division 14 of the American Psychological Association)

Publication Dates: 1987 (revision of 2nd edition); 1980; 1975

Purpose: To specify principles of good practice in the choice, development, evaluation, and use of personnel selection procedures, particularly to ensure that performance on a test (or other basis for decision) is related to performance on a job or other measures of job success. The *Principles* are intended to be consistent with the APA *Standards.*

Scope: Division 14 Principles address issues involving use and evaluation of employee selection, placement, and promotion decisions and procedures.

Intended Users: Those conducting research on selection, applying and using selection procedures, or managing validation efforts.

Topics Covered: Job analysis, criterion validation including the choice of criterion and the adequacy of research methodology, differential prediction, content validation including procedures for identifying the content domain of the job, construct validation, and validity generalization, including conditions when it is appropriate.

Reference: Society for Industrial and Organizational Psychology, *Principles for the validation and use of personnel selection procedures* (third edition) (College Park, MD: 1987).

Copies Available From: The Society for Industrial & Organizational Psychology, Inc., Department of Psychology, University of Maryland, College Park, MD 20742

"Code of Fair Testing Practices in Education"

Short title: The Code

Developed by: The Joint Committee on Testing Practices, a cooperation of American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education

Cosponsored by: American Association for Counseling and Development/Association for Measurement and Evaluation in Counseling and Development, and the American Speech-Language-Hearing Association

Endorsed by: Educational Testing Service, The College Board, American College Testing Program, CTB McGraw Hill, the Psychological Corporation and the Riverside Publishing Company

Publication date: 1988

Purpose: The Code states professional test developers' and users' obligations to test takers. It is consistent with relevant parts of the *Standards for Educational and Psychological Testing* (1985). Endorsers commit themselves to safeguarding the rights of test takers by following its principles.

Scope: The Code applies broadly to the use of tests in education (admissions, educational assessment, educational diagnosis, and student placement), but not to tests made by individual teachers for use in their own classrooms. It "is not designed to cover employment testing, licensure or certification testing, or other types of testing."

Intended Users: The general public, test takers and their parents or guardians; and professional test developers and users, particularly commercial test publishers.

Topics covered: The Code states test developers' and users' obligations in developing or selecting tests, in interpreting scores, in striving for fairness, and in informing test takers. The first of any guidelines or standards to address the "truth in testing" issues, the Code states merely that test takers should be informed of their rights.

Some examples with respect to validation studies and "truth in testing" issues are:

Test developers should "Investigate the performance of test takers of different races, gender, and ethnic backgrounds when samples of sufficient size are available" and "Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors" (C-15).

Test users should "Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use" (B–11).

Professionals that control the tests and test scores should "Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets. . . . " (D–21).

Reference: Joint Committee on Testing Practices, *Code of FairTesting Practices in Education* (Washington, DC: American Psychological Association, 1988).

Copies Available From: National Council on Measurement in Education, 1230 Seventeenth Street, NW, Washington, DC 20036; or Joint Committee on Testing Practices, American Psychological Association, 1200 7th Street, NW, Washington, D.C. 20036

ETS Standards for Quality and Fairness

Short Title: ETS Standards

Developed By: Educational Testing Service

Publication Date: 1987, 1981

Precursor: Principles, Policies and Procedural Guidelines Regarding ETS Products and Services (1979)

Purpose: The ETS Standards reflect the APA Principles tailored to the needs of this large test developer.

Scope: ETS educational and employment testing practices, programs, or services.

Intended Users: ETS professionals who must exercise professional judgment in their work.

Topics Covered: Accountability to test takers, program sponsors, professional associations, ETS founders, and the public; confidentiality of test scores and other data; technical quality of tests having to do with test development procedures, validity, test administration, and score interpretation; the promotion of fair and appropriate test use, proper interpretation of test results, and discouragement or elimination of misuse; and public understanding of testing, measurement, and related educational issues.

Reference: Educational Testing Service, *ETS Standards for Quality and Fairness*, (Princeton, NJ: 1987).

Copies Available From: Educational Testing Service, Rosedale Road, Princeton, NJ 08541–0001

"Guidelines on the Uses of College Board Test Scores and Related Data"

Short Title: College Board Guidelines

Developed By: The College Board

Publication Date: 1988

Purpose: To promote educationally sound use of college entrance test scores, examination grades, and related information; to highlight proper and beneficial uses of test scores and related data; and to caution against uses that are inappropriate.

Scope: Use of educational test scores and related data provided by the College Board.

Intended Users: The College Board; schools, colleges, universities, scholarship agencies, and other organizations using College Board Test scores and related information; counselors, college recruiting officials, and school admissions personnel.

Topics Covered: The limitations of testing, the importance of using test results in conjunction with other information, the importance of validity studies conducted by test users, fairness, problems of "overusing" test results either by interpreting scores too broadly or too precisely, avoiding the misuses of test scores, appropriate uses of aggregate scores (i.e., classroom or school averages), and the rights of test takers to privacy.

Reference: The College Board, "Guidelines on the Uses of College Board Test Scores and Related Data" (New York: 1988).

Copies Available From: College Board National Office, 45 Columbus Avenue, New York, NY 10023–6992

"Guidelines for Proper Use of NTE Tests"

Short Title: NTE Guidelines

Developed By: The NTE Policy Council and ETS. The Policy Council for the National Teachers Exams represents State departments of education and school districts that use the tests, user and nonuser teacher training institutions, and practicing classroom teachers.

Publication Dates: 1985, 1979, 1974, 1971

Purpose: To help ensure correct and appropriate use of NTE tests

Scope: Testing for a variety of purposes related to the teaching profession. They include admissions to teacher preparation programs, requirements for college graduation, program evaluation initial certification, renewal/recertification, course equivalents, and identification of candidates for employment selection.

Intended Users: State agencies responsible for credentialing, teachers; school districts, colleges, and universities; and State governing boards for public higher education.

Topics Covered: These Guidelines encourage users to rely upon multiple criteria in making selections or certifications; publicly promulgate these criteria; validate tests locally by complying with professional and legal standards as when they require job analyses, by ensuring that test content is appropriate for teacher-training programs and job requirements and by using an explicit process for and appropriately justifying cut scores; avoid overinterpreting test scores such as in evaluating experienced teachers; and avoid rank ordering candidates and other misuses of test scores.

Reference: Educational Testing Service, "Guidelines for Proper Use of NTE Tests," (Princeton, NJ: 1985).

Copies Available From: Educational Testing Service, Rosedale Road, Princeton, NJ 08541–0001.

Appendix B

Major Legislation and Litigation Involving Testing

I. Employment Testing

A. Evolution of Standards for the Use of Tests in Employment

Griggs v. Duke Power Company, 401 U.S. 424 (1971).

Thirteen black employees brought suit against their public utility employer alleging employment discrimination. Company policy required a high school diploma and minimum test scores as prerequisites for employment in, or transfer to, jobs at the plant. This policy disqualified blacks at a rate disproportionately higher than whites. Lower courts found no showing by plaintiffs that the defendant employer had adopted the diploma and test requirements with a discriminatory purpose. The Supreme Court, however, struck down the use of the criteria, reasoning that it was unrelated to job performance. The Court did not require plaintiffs to show the employer established the criteria with any *discriminatory intent*, finding that employment practices which are *discriminatory in their consequences* violated section 703(a)(2) of Title VII. Under the "disparate impact" standard, unless the employment practice (e.g., a test) can be shown to be a valid predictor of job success or can be otherwise shown to be a business necessity (i.e., "demonstrates a manifest relationship to the employment in question" *Id.* at 432), the practice or criteria (in this case a high school diploma and minimum test scores) is considered a violation of Title VII § 703(a)(2).

This case established the concept of disparate impact which, when clarified and extended by later cases, became the three-pronged analysis now commonly used in both employment and education testing litigation: the plaintiff must first establish a prima facie case of discrimination; then the defendant employer must demonstrate that the test is a business necessity, i.e., he must show that the test bears "a demonstrable relationship to successful performance of the jobs for which it [is] used" (*Id.* at 4310); finally, the plaintiff may prevail by offering either an equally effective alternative practice that has a less discriminatory impact or proof that the legitimate practices are a pretext for discrimination.

Albemarle Paper Company v. Moody., 422 U.S. 405 (1975).

Former black employees alleged that, among other things, employer Albemarle's testing program had a disproportionate adverse impact on blacks, was not shown to be related to job performance, and selected in a racial pattern significantly different from that of the pool of applicants. The Supreme Court overruled lower court findings that the test was proven to be job related by validation studies. It declared that employers must use professionally accepted validation methods to demonstrate that the test is "predictive of or significantly correlated with important elements of work behavior which comprise or are relevant to the job or jobs for which candidates are being evaluated." The Court asserted that this standard is required by its holding in *Griggs* and by the Equal Employment Opportunity Commission's (EEOC) Guidelines for judging validity and job relatedness.[1]

Watson v. Fort Worth Bank, 108 S.Ct. 2777 (1988)(Plurality opinion).

A black bank teller, rejected in favor of white applicants for promotion to a supervisory position at the bank, alleged that the bank's policy of using the *subjective judgment* of supervisors acquainted with job requirements and candidates, rather than precise and formal selection criteria, constituted discrimination in violation of Title VII of the Civil Rights Act of 1964.[2] The Supreme Court held for the plaintiff by extending the use of the disparate impact standard to subjective judgments of performance or potential, such as interviews or performance appraisals.

---

[1] The case relied upon the "Guidelines on Employee Selection Procedures," published in 1970. The current "Uniform Guidelines on Employee Selection Procedures" had not yet been published and were issued jointly by EEOC and other agencies. See 29 C.F.R. § 1607.

Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989).

This case did not involve testing. However, it affected the standard of proof for disparate impact ananlysis. The disparate impact standard applies to all elements of the hiring or promotion process, including any tests.

A group of Eskimo and Asian workers had filed a class action suit against their previous employers, two Alaskan canneries, alleging employment discrimination because they had been channeled into lower paid, unskilled jobs while the more desirable positions went to whites. The Supreme Court ruled that the plaintiff's statistical evidence was inadequate to require the employers to meet their burden of proving the "business necessity" of their employment policies.

This majority opinion affirmed the scheme for shifting the burden of proof from plaintiff to defendant and then back to plaintiff and for the evidentiary standards first laid out in *Watson* and later addressed in the Civil Rights Act of 1991.

Civil Rights Act of 1991

The Civil Rights Act of 1991 restored the burden of proof and standards (e.g., the concepts of "business necessity" and job related) prevailing before the 1989 *Wards Cove v. Atonio* decision.[3] Also, it clarifies that the complaining party must demonstrate a disparate impact for each particular challenged employment practice, except if he demonstrates that the elements of an employer's decisionmaking process are not capable of separation for analysis, he may analyze it as one employment practice.

Because the Civil Rights Act of 1991 overrides most of the effects of *Wards Cove*, the primary recent change with respect to testing is *Watson's* extension of validation procedures to subjective measures of performance. Subjective criteria for selection, a viable alternative to tests in the past, will now require justification or validation when they show adverse impact. Employers' selection procedures may change. More generally, the passage of the Civil Rights Act of 1991 strengthens the deterents against discrimination and provides better protection for those who suffer employment discrimination.

B. Employment Testing in Federal Agencies

Luevano v. Campbell, 93 F.R.D. 68 (D.D.C. 1981).

A class action employment discrimination suit alleged that the PACE exam, a test developed and administered by the U.S. Office of Personnel Management (OPM) and used to hire professional-level applicants for Federal jobs, had an adverse impact on blacks and Hispanics. In a consent decree OPM agreed to phase out the PACE over a 3-year period and henceforth to administer separate examinations for most of the current PACE job categories.[4]

OPM replaced the PACE with procedures requiring applicants to complete the Standard Form 171 (SF-171) giving detailed information about past jobs and educational curriculum and

---

[2] 42 U.S.C. §§ 2000e *et seq.* provides in pertinent part that "it shall be an unlawful employment practice for an employer (1) to fail or refuse to hire or to discharge any individual, or otherwise to discriminate against any individual with respect to his compensation, terms, conditions, or privileges of employment, because of such individual's race, color, religion, sex, or national origin; or (2) to limit, segregate, or classify his employees or applicants for employment in any way which would deprive or tend to deprive any individual of employment opportunities or otherwise adversely affect his status as an employee, because of such individual's race, color, religion, sex, or national origin."

[3] Wards Cove Packing Co. v. Atonio, 490 U.S. 642 (1989). In this case, the Court shifted the burden of proof applicable in disparate impact cases from the defendant to the plaintiff. See A Report of the U.S. Commission on Civil Rights *The Civil Rights Act of 1990* (July 1990) for an analysis of the *Wards Cove* decision.

[4] *See also* 26 A.L.R. Fed. 13.

BEST COPY AVAILABLE

compelling the hiring agency's personnel officials to rate the applications according to educational coursework and experience. The required efforts of both applicants and agency staff made these procedures time consuming and resulted in delays in hiring. Finally, in 1990, a new test, known as the Administrative Careers with America (ACWA), was implemented. The ACWA is expected to streamline the hiring process and have less adverse impact than the PACE.

Palmer v. Shultz, 616 F. Supp. 1540 (D.D.C. 1985), *rev'd and remanded* 815 F.2d 84 (D.C. Cir. 1987).

A female foreign service officer alleged that the State Department engaged in an unlawful discriminatory employment practice when using the Foreign Service Exam (FSE) to assign women to areas of job specialization. The plaintiff alleged the test had a disparate *impact* on females resulting in their disproportionate *overassignment* to consular positions and *underassignment* to political positions during 1976–1983. The court concluded that no significant statistical disparity in job assignments was present because appointments were tied to test scores and women preferred consular positions. On remand, however, the court found that the preferences of women applicants were unknown or irrelevant to the defendant and therefore did not excuse the overinclusion of women in consular positions. The court also found that the FSE was not job related and had a disparate impact on women who scored lower than men on the political portion of the test.[5]

The district court found that the plaintiffs had failed to prove their case by a preponderance of the evidence and entered judgment for the defendant. Plaintiff appealed.[6] On appeal the court reversed the lower court's holding, finding its decision erroneous in a number of instances. The appeals court remanded the case back to the lower court with instructions to find additional facts (and the appropriate statistical analysis to use to do so)[7] before determining liability under Title VII. With this mandate, the district court on remand found that plaintiff's statistics demonstrated a violation of Title VII and that the defendant had failed to rebut them or to show that the specialized portion of the written examination was job related (662 F. Supp. at 1571). Judgment was entered for the plaintiffs on all claims except those relating to discrimination in promotions.

The State Department canceled the 1989 test results for nearly 15,000 applicants after this ruling and began searching for a new, bias-free grading methodology.[8]

II. Testing in Education

A. Elementary Schools

Hobsen v. Hansen, 269 F. Supp. 401 (D.D.C. 1967) *cert. dismissed*, 393 U.S. 801 (1967).

This was an allegation that a school system's policy of grouping students by ability using scores on group administered aptitude tests violated the 5th and 14th amendments to the Federal Constitution. Judge Skelly Wright declared that the system discriminated unconstitutionally because the standardized tests produced inaccurate and misleading scores. He ordered the Washington, D.C., public school system to abolish its educational tracking system and provide the court with a plan of "pupil assignments complying with the principles announced in the court's opinion. . . ."[9] The court, however, cautioned that "not all classifications resulting in disparity are unconstitutional. If a classification is reasonably related to the purposes of the governmental activity involved and is rationally carried out, the fact that persons are thereby treated differently does not necessarily offend." The court, thus, condemned the use of rigid, poorly conceived classification practices that damaged the educational opportunities of minority children. It did not prohibit "ability grouping" per se, but only as practiced in this District.[10]

---

[5] See 661 F. Supp. at 1571, n. 34 & 35.

[6] 815 F.2d 84 (D.C.Cir. 1987).

[7] Seger v. Smith, 738 F.2d at 1249, 1282 (1984).

[8] See John Purnell, "Bias ruling shelves Foreign Service test results," *The Washington Times*, Mar. 9, 1989; and Enrique J. Gonzales, "Lawmaker charges discrimination at State," *The Washington Times*, Oct. 13, 1989.

[9] 269 F. Supp. at 517.

Larry P. v. Riles, 495 F. Supp. 926 (N.D. Cal, 1979), *aff'd in part and rev'd in part*, 793 F.2d 969 (9th Cir. 1984).

Black students, in a suit brought by their parents on their behalf, alleged that the school system's use of IQ tests, without establishing the validity of such a test as an educational necessity for placing children in classes for the educable mentally retarded, violated regulations issued under Title VI of the Civil Rights Act of 1964.[11] The parents complained that their children were placed in classrooms for the educably mentally retarded at over twice the rate of white children and that this stigmatized them and did irreparable harm to their educational advancement. District Court Judge Robert Peckham issued an order, which was affirmed by the court of appeals, prohibiting the school district from using IQ tests for placing black children in classes for the educably mentally retarded. In 1986, the order was expanded, banning IQ testing throughout the State of California for evaluation, admission, and placement of black schoolchildren with learning disabilities (including the mentally retarded). The expanded ban was challenged in *Crawford v. Honig*, below.

Parents in Action on Special Education (PASE) v. Hannon, 506 F. Supp. 831 (D.C. Ill. 1980).

The case alleged that IQ tests administered by the Chicago board of education were culturally biased against black children and that the use of such tests violated the equal protection clause and Title VI of the Civil Rights Act of 1964.[12] The court held that the challenged test, taken as a whole and in conjunction with the statutorily mandated other criteria for determining an appropriate educational program for a child, did not discriminate in violation of the Constitution or statute.

Although the issues, and in many instances the evidence and expert witnesses, were the same as in *Larry P.*, the decision was opposite. After examining hundreds of test items, Judge Grady concluded that except for a few items, the IQ tests were fair, i.e., they were as useful for a black as for a white child in making educational decisions. (Although the IQ tests were vindicated, they were later banned as part of a desegregation settlement.)[13]

Crawford v. Honig, No. 89-0014-RFP (N.D. Cal. May 10, 1988).

Parents of black students alleged that California may not refuse to provide IQ testing to their children when they request it, and the test is available to other children including whites. The parents believed an IQ test would prove the children do not belong in special education classes. California refused to administer the test pursuant to a Federal court order issued in *Larry P. v. Riles*. The case challenged the 1986 expansion of the ban on testing.

---

[10] 269 F. Supp. at 511, *citing to* Justice Burton's opinion for the court and Justice Frankfurter's dissent in Morey v. Doud, 354 U.S. 457 (1957). The classification which fell short of the requirements of the rational basis test was "ability grouping" *as administered and practiced* by the District's school system. The court in *Hobsen* made it clear that "the concept of ability grouping . . . can be reasonably related to the purposes of public education." 269 F. Supp. at 512.

[11] 42 U.S.C. § 2000d *et seq.* provides that "No person in the United States shall, on the grounds of race, color, or national origin, be excluded from participation in, be denied benefits of, or be subjected to, discrimination under any program that receives federal financial assistance." *See also* 34 C.F.R. 100.3(b)(2) for regulations issued under this statutory mandate. The plaintiffs also relied on the Rehabilitation Act of 1973. 29 U.S.C. § 794 et. seq. (1988).

[12] *See supra* note 18.

[13] *See also*, 44 A.L.R. Fed. 148.

In September 1992, U.S. District Court Judge Robert F. Peckham lifted the testing ban.[14]

Montgomery v. Starkville Muni. Separate School Dist., 854 F.2d 127 (5th Cir. 1988).

Intervenors in a school desegregation suit sought injunctive relief on the basis that the school district's use of achievement groupings in certain subjects and grades constituted a dual system of education with a disproportionate number of white children in the more advanced achievement groups. They charged that this result was contrary to the Supreme Court holding in *Brown v. Board of Education*, 347 U.S. 483 (1954). On appeal, the court for the fifth circuit concluded that the school district's achievement grouping was properly employed for the purpose of assisting students in their ability to learn.

This case has become the model guiding the Department of Education's policy in monitoring school districts. The Department generally considers whether achievement tests are used instead of ability tests, the amount that teacher judgments contribute to decisions about ability grouping, and whether there are some subjects which are not grouped by ability.

Quarles v. Oxford Muni. Separate School Dist. 868 F.2d 750 (5th Cir. 1989).

The Oxford school district used a limited form of achievement grouping that the Department of Education's Office for Civil Rights had reviewed, modified, and approved. The district court found that the grouping system was educationally sound in theory and practice. The court also found that students were not locked into a given group and could move about between levels under certain circumstances.

On appeal the appellant argued that the school district's achievement grouping discriminated against black students on the basis of race in violation of the 14th amendment and Title VI. The court, however, disagreed and pointed out that "[a]chievement or ability grouping has been recognized as an acceptable and commonly used instruction method."[15] The court found that the district's grouping of grades 3–8 in language and math based on the Stanford Achievement Test (SAT) neither intends nor produces a "significant racial impact upon the makeup of the classroom. . . ."[16]

B. Minimum Competency Testing For Students and Teachers

Debra P. v. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), *aff'd in part and vacated*, 644 F. 2d 397 (5th Cir. 1981), reh'g denied, 654 F.2d 1079 (5th Cir. 1981); on remand 564 F. Supp 177 (M.D. Fla. 1983), *aff'd*, 730 F.2d 1405 (11th Cir. 1984).

A class action suit alleged that Florida's recently implemented functional literacy exam denied black students due process and equal protection of the law. The students were denied diplomas after failing the test which had a passing score aimed at a minimum competency level. Evidence indicated that the minimum competency criterion would have denied diplomas to 20 percent of black, but only 2 percent of white, high school seniors.

The court issued a temporary injunction prohibiting the testing program because its abrupt implementation perpetuated the effects of past discrimination lingering in the Florida school system.[17] The injunction was dissolved 4 years later when all students taking the exam had been trained exclusively in the presently desegregated schools.

---

[14] This description is based upon conversations with staff of the Landmark Legal Foundation. The foundation represented the plaintiffs. Also, see Jean Merl, "Court Ban on IQ tests for Blacks Sparks Parents' Suit," *The Los Angeles Times*, Aug. 5, 1991; "Judge lifts ban on IQ testing," *The Washington Times*, Sept. 3, 1992; and "Judge lets California Resume IQ Testing of Black Students," *Education Daily*, Sept. 8, 1992, p. 4.

[15] 868 F.2d at 753.

[16] *Id.* at 755. The student's teachers could change the initial achievement grouping if exceptional progress or lack of progress indicates that movement may be proper. *Id.*

An appellate court held that the test was valid if its contents were actually taught in the schools. Furthermore, the test was appropriate if the State could prove either that any racially discriminatory impact was not due to the present effects of past intentional discrimination or that the use of the test would remedy such present effects. After additional evidence was presented the district court found that the test was instructionally valid and that the present effects of past discrimination did not cause the disproportionate failure rate of black students.

Anderson v. Banks, 520 F. Supp. 472 (S.D. Ga. 1981).

Plaintiffs alleged that the county school district's requirement that students pass a reading and math test at a ninth grade level as a precondition to receiving a diploma was racially discriminatory in view of the fact that some students had attended substandard segregated schools under the previous system of dual education. The court held that the testing requirement *as applied* violated Title VI of the Civil Rights Act of 1964 and the Equal Educational Opportunities Act.[18] The decision rested on the historical fact that racially segregated classes, produced by a discriminatory tracking system, placed black children in lower ability classes than white students with identical test scores. The court suggested that the school could return to the test and diploma requirement once all of the students formerly segregated had passed through the system.

LULAC v. United States, 793 F.2d 636 (5th Cir. 1986).

An organization representing students challenged a requirement that students must pass a skills test to enroll in more than 6 hours of professional education at any State college or university (thereby precluding them from becoming teachers with degrees from Texas public institutions), on the basis that such a requirement violated Title VI of the Civil Rights Act of 1964.[19] The district court granted a preliminary injunction permitting otherwise qualified students to enroll in classes. On appeal the injunction was dissolved, the court holding that the lower court abused its discretion in failing to assess evidence that the skills test was a bona fide occupational qualification. The Fifth Circuit Court of Appeals held that the civil rights claim could be proven through a disparate impact analysis only if the challenged test was not a reasonable measure of bona fide education requirements.

C. Higher Education: Admissions and Scholarships

Sharif by Salahiddin v. New York State Educ. Dept., 709 F. Supp. 345 (S.D.N.Y. 1989).

Female students alleged that New York State's practice of sole reliance upon SAT scores to award scholarships disparately impacts them in violation of Title IX of the Education Amendments of 1972[20] and the equal protection clause of the 14th amendment. The defendants acknowledged that the SAT underpredicts academic performance for females as compared to males. The court determined that in selecting those to receive scholarships, the New York State Education Department's intent was to reward high school achievement. It, therefore, prohibited the State's current practice and added that the best currently available alternative would combine grades and SATs, but that other alternatives, including a statewide achievement test, could be developed in the future.

---

[17] 42 U.S.C. § 2000d and 20 U.S.C. § 1703.

[18] *See* 42 U.S.C. § 2000d *et seq.* and 20 U.S.C. 1703(b).

[19] *See supra* note 18.

[20] 42 U.S.C. § 2000d–7 provides in pertinent part that "(a) No person in the United States shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity receiving any federal assistance."

III. Test Construction Issues—Out of Court Settlements

Golden Rule Insurance Co. v. Washburn, 1984, (No. 419-76, Ill. 7th Jud. Cir.).

This was an allegation that the Illinois insurance test required for State licensure was not sufficiently related to the knowledge, skills, and abilities needed by insurance agents, and that these tests intentionally discriminated against test takers on the basis of race. The parties reached an out of court settlement whereby the test developer—Educational Testing Service (ETS)—agreed to use the "Golden Rule procedure" in assembling new test forms. This procedure reduces the number of test questions or items that are more difficult for blacks than for whites. ETS has since concluded that this settlement was a mistake (Anrig, 1987).[21]

Allen v. Alabama State Bd. of Educ., 636 F. Supp. 64 (M.D. Ala. 1985), *rev'd* 816 F.2d 575 (11th Cir. 1987).

Teacher candidates sued the State to enjoin the use of a basic competency test required for certification, which blacks had failed in disproportionate numbers. A settlement provided for reinstatement of several hundred failed teacher candidates. It also mandated that any future test developed would have to abide by a variant of the "Golden Rule procedure" (i.e., use as preferred items those on which passing rates between blacks and whites did not differ by more than 5 percent, with fallback items permitting a difference of 10 percent and no more than 10 percent of the items having up to a 15 percent differential).

---

[21] Anrig, Gregory R., "ETS on 'Golden Rule'," *Educational Measurement: Issues and Practice*, Fall 1987: 24–27.

*U.S. G.P.O.:1993-350-410:80322